A

Abrasion

► The Tribology of Dental Materials

Abrasive Machining Processes

► Grinding Processes

Abrasive Wear of Gears

► Wear in Gears

ABS – Air Bearing Surface

► ABS Designs

ABS Designs

David B. Bogy[1], Du Chen[2]
[1]Computer Mechanics Laboratory, Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA, USA
[2]Western Digital Corporation, San Jose, CA, USA

Synonyms

ABS – air bearing surface

Definition

The air bearing surface in a hard disk drive is the surface of the slider facing the disk and is aerodynamically lubricated.

Scientific Fundamentals

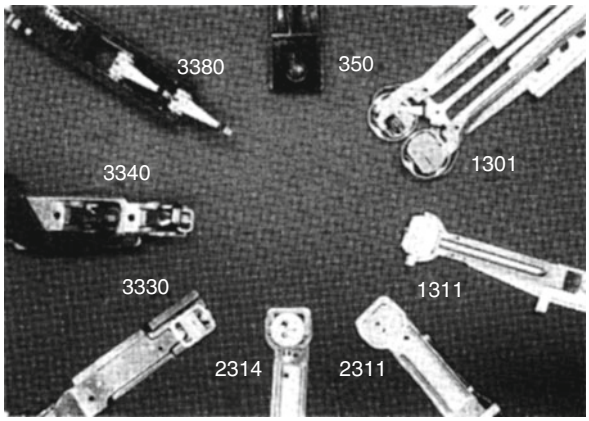
Modified Compressible Reynolds Equation

Lubrication theory and its application to journal bearings and the like was one of the primary means of the Industrial Revolution in the nineteenth century. This theory is based on the fluid mechanics of incompressible fluids (oil, etc.) in "thin" regions. The lubrication equation, commonly called the Reynolds equation, is derived from the incompressible form of the Navier–Stokes equations of fluid mechanics. For application to air or other gas bearings of hard disk drives (HDD), the lubrication equation for compressible fluids is required. This was recognized and developed by Gross (1959), and also in his subsequent book (Gross 1962), which are must reading for students or engineers entering the field of HDD air bearing design. It was recognized in the late 1950s that if the spacing in the gas bearing becomes low in comparison to the mean free path of the gas molecules, i.e., the gas is said to be rarefied, or the Knudson number $K_n = \lambda/h \rightarrow 1$ or even larger, where λ is the mean free path distance and h is a characteristic spacing, then the usual non-slip boundary conditions of fluid mechanics would no longer hold.

Slip models are the bases of slip-corrected compressible Reynolds equations for rarefied gas lubrication in modern HDDs. A compressible Reynolds equation is derived from the Navier–Stokes equations with velocity boundary conditions, the conservation of the mass flow rates, the equation of state of the compressible flow and the velocity boundary condition. In hard disk drives the disk's rotation velocity is low subsonic and the gas film thickness is much smaller than the slider's width and length. Usually it is assumed that the flow is isothermal, the pressure field is uniform in the film thickness direction, the inertial effects are negligible, and the viscosity change due to the position and velocity is also negligible. The slip model prescribes

Q.J. Wang, Y.-W. Chung (eds.), *Encyclopedia of Tribology*, DOI 10.1007/978-0-387-92897-5,

the velocity boundary condition due to rarefaction, which usually depends on the velocity profile near the wall, the Knudsen number, the surface accommodation coefficient, the flow velocity gradient or pressure gradient, or shear stress at the wall.



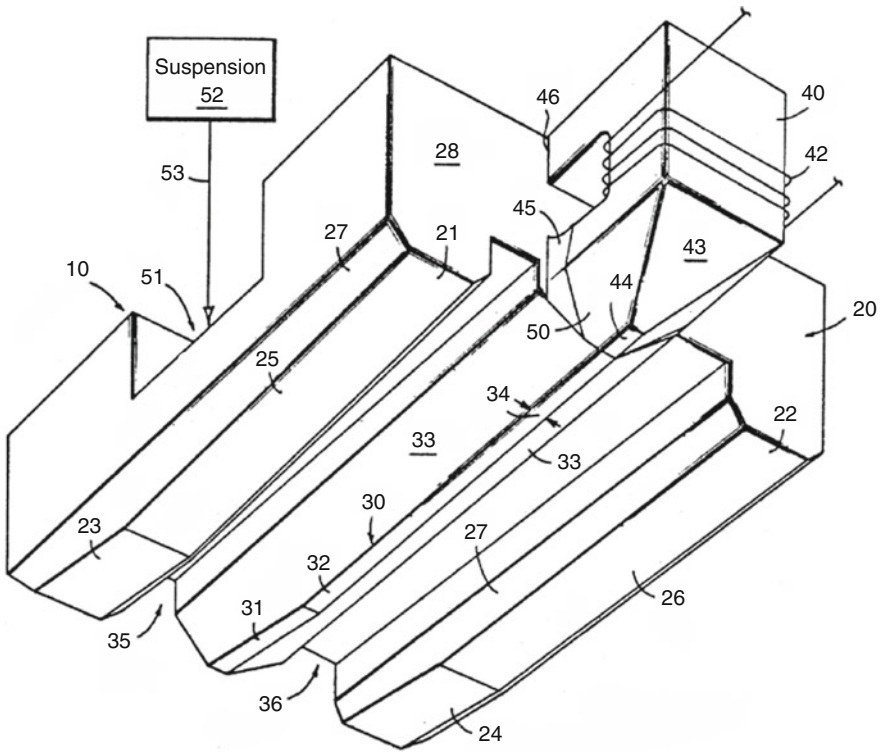
ABS Designs, Fig. 1 Genealogy of air-bearing sliders with their magnetic heads and support structures from 1957 to 1981 (Gross 1984)

In general the obtained slip-corrected, time-dependent Reynolds equations have the form,

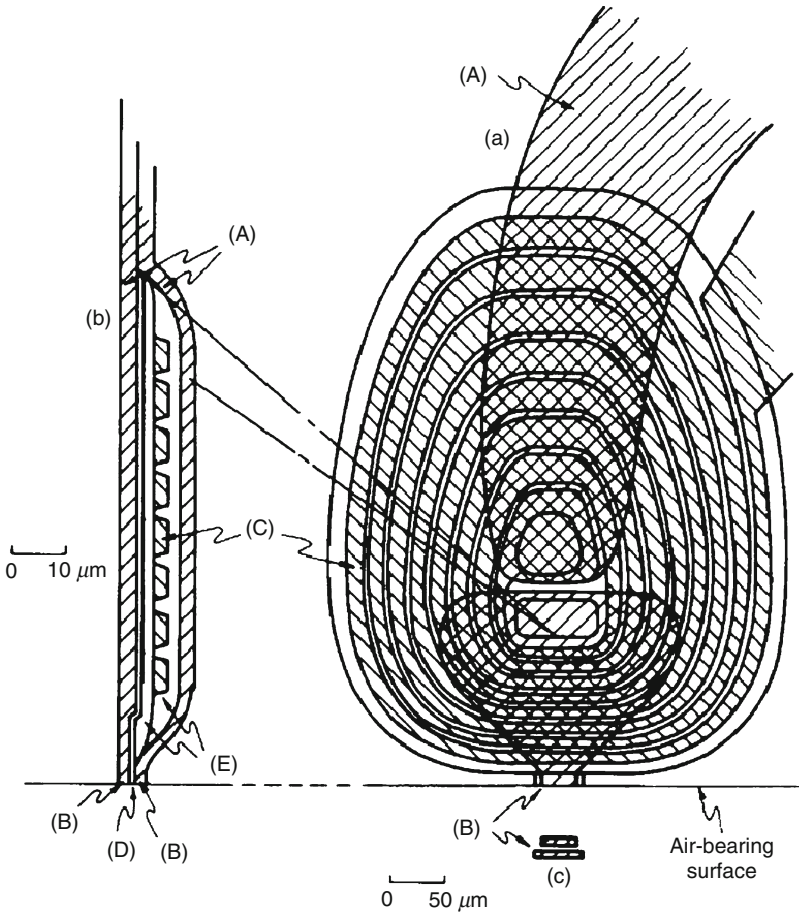
$$\nabla \cdot [(\bar{Q}_P p h^3 \nabla p) - 6V\mu p h) = 12\mu \frac{\partial (ph)}{\partial t}, \qquad (1)$$

where $\bar{Q}_P$ is the relative Poiseuille flow rate coefficient, p is the gas pressure, V is the relative velocity of the bearing surfaces, μ is the gas viscosity, and t is time. The deviation of its value from 1 represents the effect of the velocity slip on the gas flow. Its expression in terms of the Knudsen number depends on the slip model used in the derivation. The net force of the gas pressure on the bearing surface is the bearing force, which supports the working load. As indicated by the equation, the gas pressure distribution under the ABS and the spacing depend on each other, while the gas pressure is directly affected by the aerodynamic surface design. In this situation, a well-designed ABS can achieve a self-regulating spacing.

Usually a compressible slip-corrected Reynolds equation does not contain the relative Couette flow rate coefficient, provided that the surface



ABS Designs, Fig. 2 Three rail taper flat "Winchester" slider from Warner patent (Warner 1974)

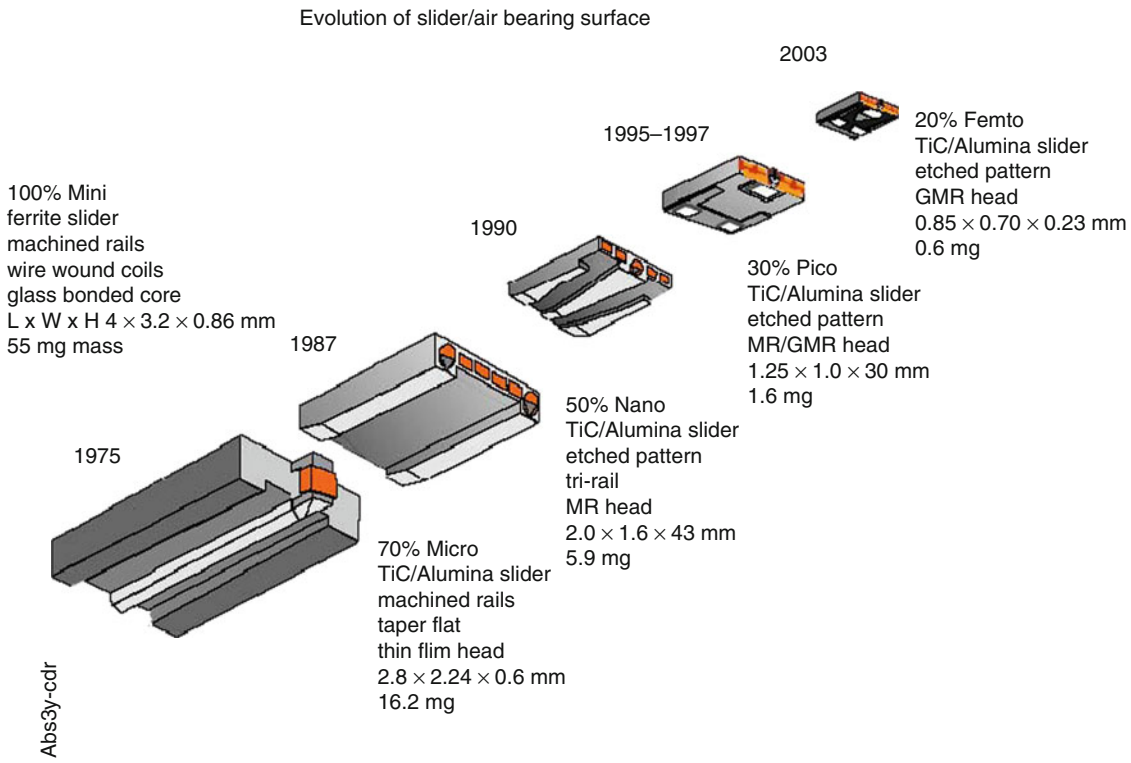


ABS Designs, Fig. 3 IBM 3370 thin-film head (Harker et al. 1981)

accommodation coefficients on both boundaries are close to each other. The surface accommodation coefficient is defined for the tangential momentum exchange of gas molecules with surfaces and it determines the velocity slip property of the boundary for slip flows. Due to the skew symmetry of Couette flow with respect to its center plane, the Couette flow rate does not depend on the velocity slip at the boundaries, provided that the slip conditions at the upper and lower boundaries are the same. On the contrary, the symmetry of Poiseuille flow with respect to the center plane results in a strong dependence of the Poiseuille flow rate on the slip conditions at the boundaries. However, slip-corrected Reynolds equations have not been widely used in the air bearing simulations since Fukui and Kaneko derived a more accurate generalized lubrication model based on the linearized Boltzmann equation (Fukui and Kaneko 1988). That model gives a special modified Reynolds equation with a flow rate $\bar{Q}_P$ as a function of the Knudsen number, which is determined by the solutions to the linearized Boltzmann's equation.

Solution Methods

The solution to the modified Reynolds equation is the basis of the ABS designs. However, the Reynolds equation is too complicated for closed-form solutions, except in very specific cases (Gross 1962). The first solutions of the air bearing equations for HDD designs used the finite differences method to solve the equations numerically (Michael 1959). Since that time various



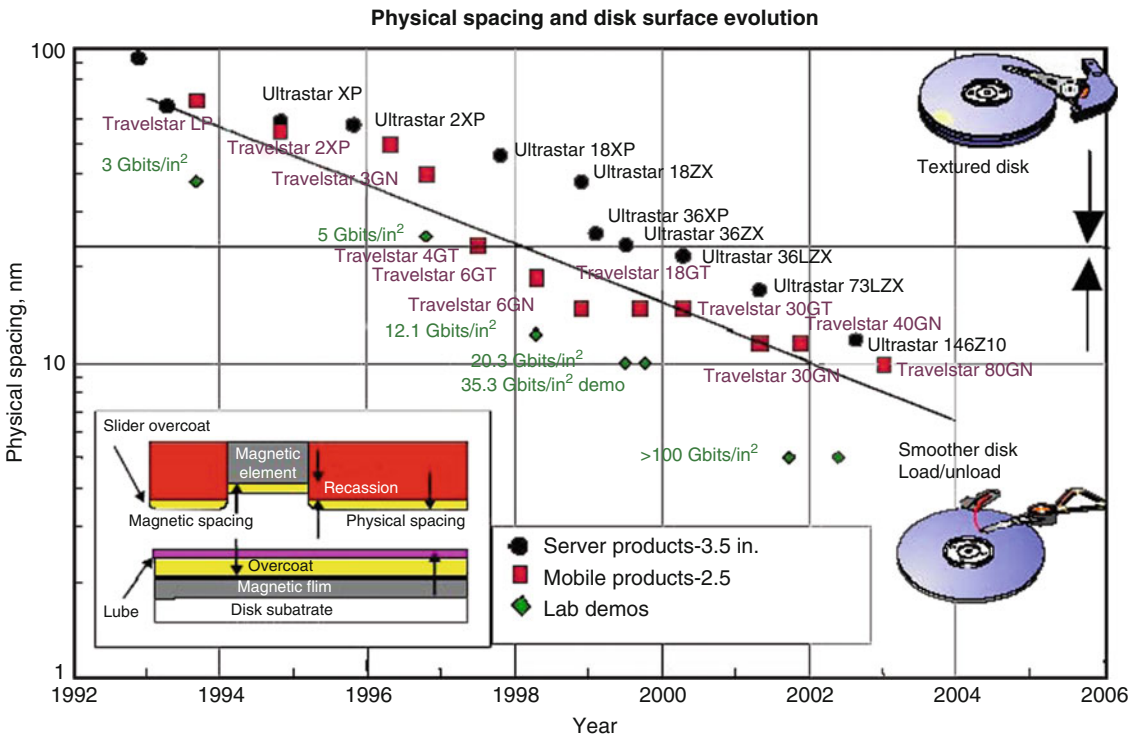
ABS Designs, Fig. 4 Evolution of slider/air bearing surface (http://www.hitachigst.com/)

schemes have been employed, including different approaches with finite differences (White and Nigam 1980), finite elements (Garcia et al. 1984), and finite volume (Lu 1997). A simulation code, CMLAIR, based on Lu's work was developed at the Computer Mechanics Laboratory in UC Berkeley. This code has been widely used by the HDD industry for many years for designing air bearings and also for simulation of such dynamical processes as shock, load-unload, response to contact, contamination studies, etc.

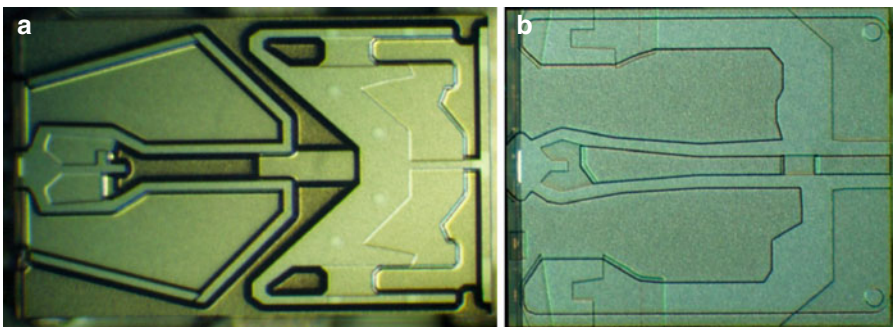
Key Applications

Historical Milestones in Air Bearing Evolution

As an important application of gas lubrication, a self-regulating air bearing provides the separation and reliability of the head-disk interface in the HDD. One boundary of the air bearing is the data disk that contains the magnetic media on which bits are written and erased, and which rotates at speeds between 5,000 and 15,000 rpm in current commercial products producing linear speeds at the read-write head up to 60 m/s. The other boundary of the air bearing is the slider, which floats above the disk with an equilibrium spacing that is determined by the interface air bearing pressure pushing the slider away from the disk and the spring loaded suspension that urges the slider toward the disk. The slider carries the reading and writing transducers usually located at the center of its trailing edge, and by design produces the spacing required by the magnetic design of the HDD. In order to increase the density of data bits along a circular track on the disk the magnetic spacing, along with other key dimensions, must be reduced. So the trend in the evolution of magnetic recording sliders in HDDs has been toward closer spacing between the slider and disk, which has been achieved by a progressive miniaturization of the slider and a corresponding reduction in the suspension load applied to the slider.



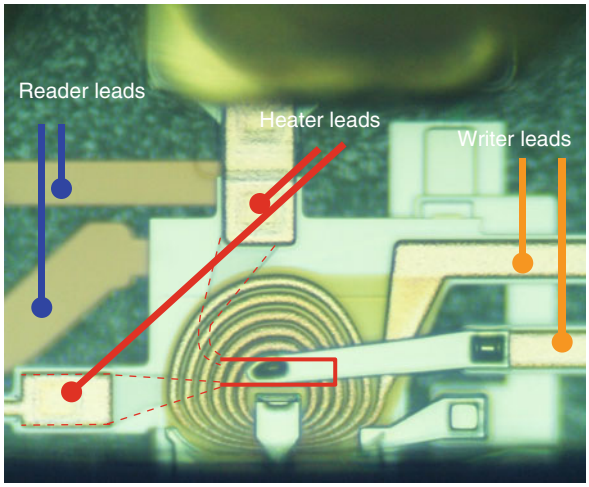
ABS Designs, Fig. 5 Reduction of the physical spacing in HDDs (http://www.hitachigst.com/) (**a**) Pemto (**b**) Femto



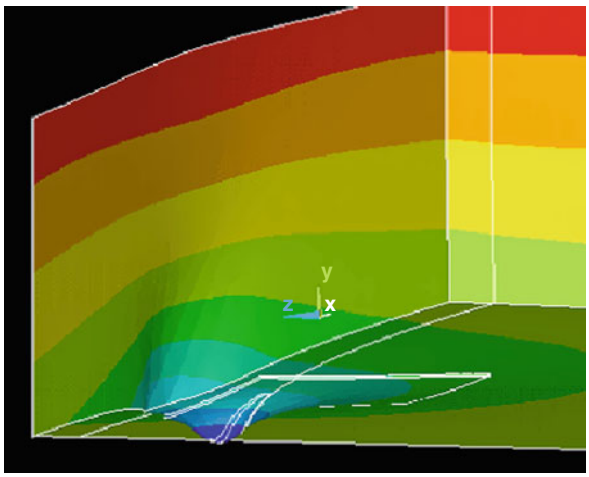
ABS Designs, Fig. 6 Examples of air bearing designs with TFC in current HDD products (transducers and heaters at *left-center*)

The evolution of the air bearing sliders from the first HDD (IBM RAMAC 350, 1957) to the first 1 GB HDD (IBM 3380, 1980) is shown in Fig. 1 (Gross 1984). As can be seen the shapes of the sliders were varied from circular disks to rectangular blocks with convex faces toward the disk and with holes or slots. The first HDD to use a self-acting slider was the IBM 1301 shipped in 1962. It was designed on the basis of the seminal three part paper published in 1959 (Gross 1959; Michael 1959; Brunner et al. 1959), which developed

lubrication theory for compressible gasses. Its flying height was about 6,350 nm. The first low mass "Winchester" three rail taper flat slider, based on the US Patent 3,823,416 (Warner 1974), was used in the IBM 3340 and shipped in 1973. Figure 2 shows a drawing of the disk facing side of that slider where it is seen that the magnetic transducer is a hand wrapped horseshoe magnetic. The flying height of this slider was 460 nm. The first thin film transducer was introduced with the IBM 3370 slider shipped in 1979, with a flying height of 330 nm. A drawing of that transducer is shown in Fig. 3. The miniaturization of air bearing sliders and the reduction in flying height has continued up to the present as data densities on hard disks have increased. Fig. 4 shows the evolution of the sliders, and Fig. 5 shows the reduction in flying height, where it is seen that the physical slider-to-disk minimum spacing was about 100 nm in 1992 and had reduced to about 12 nm in products by 2002. Figure 4 also shows an important development around 1990 when slider manufacturing changed from grinding (straight rails) to etching (any shaped rails). This advance made it possible to achieve uniform flying heights over the disk from the inner radius to the outer radius, using a swing arm actuator that causes skew of the slider's center line with respect to the tangent of the data track. Using various etch depths together with complex air bearing shapes vastly complicated air bearing design and opened the way to optimized performance taking into consideration many requirements and parameters. As seen in Fig. 4 the physical spacing was projected to reach about 5 nm in 2006, a spacing below which destabilizing forces, such as van der Waals forces and meniscus forces come into play (Wu and Bogy 2002). In order to lower the transducer further, an important development (Meyer et al. 1999) called thermal fly-height control (TFC) or dynamic flying height (DFH), has been employed in air bearing sliders since 2007. TFC employs a heating element integrated in the thin film transducer structure. When power is supplied to this element the slider thermally expands locally in a way that protrudes only a small region around the read-write transducers to move it closer to the disk. Because of the smallness of the close-approach region the destabilizing forces are minimized. Using this ingenious invention the physical spacing of the read-write elements can be moved all the way down to contact if desired while the minimum spacing of the slider otherwise remains several nanometers above the disk. Such a partial contact air bearing system is being developed for data densities greater than 1 Tbit/in^2.



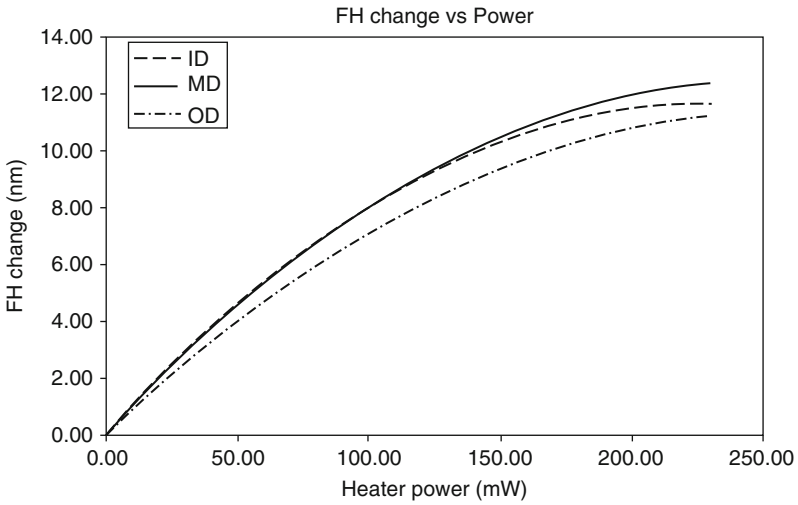
ABS Designs, Fig. 7 A typical transducer and heater design in a TFC slider



ABS Designs, Fig. 8 A finite-element calculation of the thermal deformation at the transducer location

ABS Designs for 10 Tbit/in^2

The magnetic spacing for 10 Tbit/in^2 is no more than 2.5 nm. Within this spacing there must be protective overcoats for the slider and disk, the lubricant on the disk, and the disk roughness. Therefore, there will be intermittent or continuous contact between the



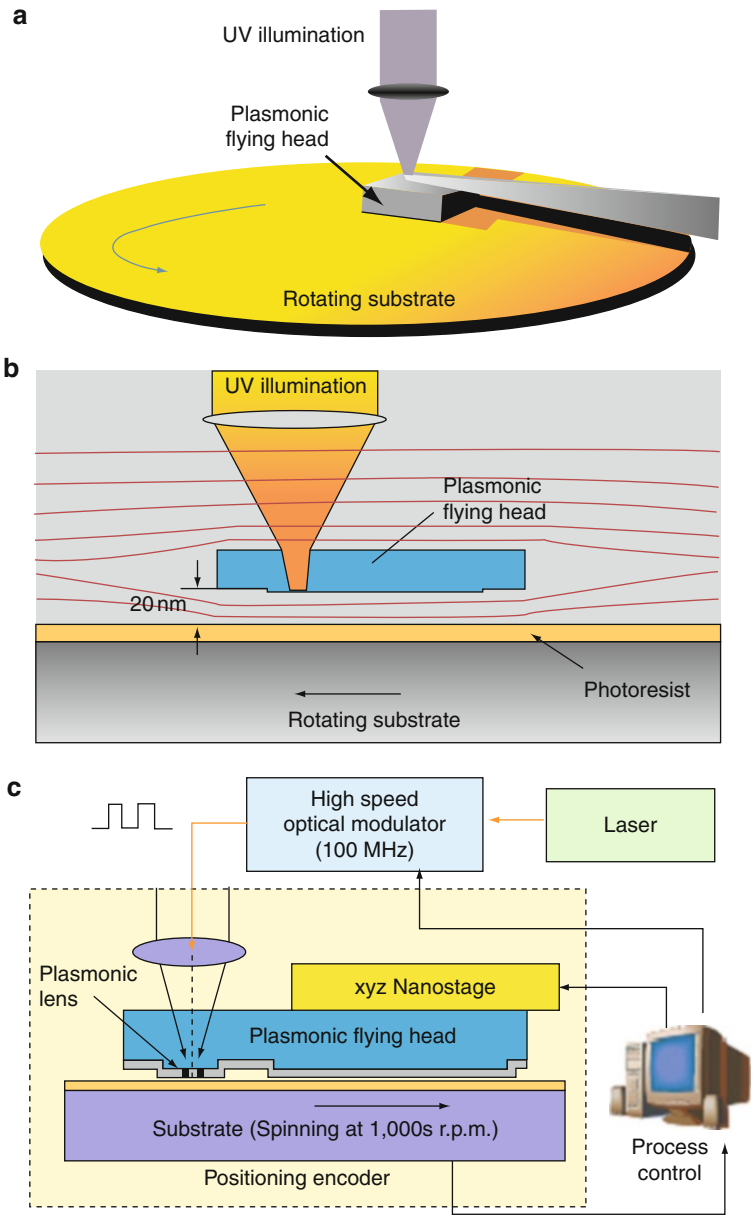
ABS Designs, Fig. 9 Power versus transducer spacing change for a typical TFC slider

read-write transducer and the disk during the reading or writing process. It appears that the only viable solution is to employ the TFC technology. The slider will fly at a nominal minimum spacing of about 8 nm when idling, and the transducer will be protruded to a light contact condition during the duty cycle. Such a system appears to be feasible based on preliminary research and development at various laboratories. Of course the critical questions related to reliability, wear, and lubricant displacement and transfer must be answered. Several designs currently employed in commercial disk drives meet these parameters, such as the designs shown in Fig. 6. Figure 7 shows a micrograph of a typical heater design. Figure 8 shows the local deformation around the transducer that is produced by the heater. Figure 9 shows a typical power versus transducer-disk spacing reduction curve for a TFC slider. These sliders fly at about 11 nm and their transducers can be brought into contact by applying about 200 mW of power to the heater. The ABS designs are required to maintain the light contact between the protrusion and the disk while damping out the vibration of the slider, so that the wear and bouncing of the head can be well controlled under an allowable limit.

There are new issues to consider in the air bearing equations, such as the contact and adhesion between the slider and disk in partial contact, the coupled heat transfer and air bearing within the slider-disk interface. Various nano-heat transfer models are being developed and tested for modifying the air bearing calculations (Chen and Bogy 2009). Also, new models for air shear at the extremely high shear rates are needed to understand its effects on the disk lubricant. If these problems are solved and a reliable head-disk interface with contact is realized then the holy grail of magnetic recording will have been achieved.

HDD Air Bearing Technology in Nano-lithography

The technology of nanometer scale flying heights of air bearing sliders has recently been employed in the area of nano-lithography. A new low-cost, high throughput approach to maskless nanolithography has been reported that uses an array of plasmonic lenses that flies above the surface to be patterned, concentrating short-wavelength surface plasmons into sub-100 nm spots (Fig. 10) (Srituravanich et al. 2008). However, these nanoscale spots are only formed in the near field, which makes it very difficult to scan the array above the surface at high speed. To overcome this problem a self-spacing air bearing was designed that can fly the array just 20 nm above the disk that is spinning at speeds between 4 and 12 m/s. The ABS of the slider embedding the plasmonic flying head is designed to meet this special mechanical requirement.



ABS Designs, Fig. 10 High-throughput maskless nanolithography using plasmonic lens arrays (Srituravanich et al. 2008)

References

R.K. Brunner, J.M. Harker, K.E. Haughton, A.G. Osterlund, A gas film lubrication study: part III, experimental investigation of pivoted slider bearings. IBM J. Res. Dev. **3**, 260 (1959)

D. Chen, D.B. Bogy, A phenomenological heat transfer model for the molecular gas lubrication system in hard disk drives. J. Appl. Phys. **105**, 084303 (2009)

S. Fukui, R. Kaneko, Analysis of ultra-thin gas film lubrication based on linearized Boltzmann equation: first report-derivation of a generalized lubrication equation including thermal creep flow. ASME J. Tribol. **110**, 253 (1988)

S.C. Garcia, D.B. Bogy, F.E. Talke, Use of upwind finite element scheme for air bearing calculations. ASME/ASLE Proceedings SP-16, 90 (1984)

W.A. Gross, A gas film lubrication study: part I, some theoretical analyses of slider bearings. IBM J. Res. Dev. **3**, 237 (1959)

W.A. Gross, *Gas Film Lubrication* (Wiley, New York, 1962)

W.A. Gross, Origins and early development of air-bearing magnetic heads for disk-file digital storage systems. Tribology and Mechanics of Magnetic Storage Systems. ASLE SP-16, 63 (1984)

J.M. Harker, D.W. Brede, R.E. Pattison, G.R. Santana, L.G. Taft, A quarter century of disk file innovation. IBM J. Res. Dev. **25**, 667 (1981)

S. Lu, Numerical simulation of slider air bearings, PhD Dissertation, Graduate Division, UC Berkeley, 1997
D.W. Meyer, P.E. Kupinski, J.C. Liu, Slider with temperature responsive transducer position, US Patent 5,991,113, 1999
W.A. Michael, A gas film lubrication study: part II, numerical solution of the Reynolds equation for finite slider bearings. IBM J. Res. Dev. **3**, 256 (1959)
W. Srituravanich, L. Pan, Y. Wang, C. Sun, D.B. Bogy, X. Zhang, Flying plasmonic lens in the near field for high-speed nanolithography. Nat. Nanotechnol. **3**, 733 (2008)
M.W. Warner, Flying magnetic transducer assembly having three rails, US Patent 3,823,416, 1974
J.W. White, A. Nigam, A factored implicit scheme for the numerical solution of the Reynolds equation at very low spacing. Trans. ASME J. Lubr. Tech. **102**, 80 (1980)
L. Wu, D.B. Bogy, Effect of the intermolecular forces on the flying attitude of Sub-5 NM flying height air bearing sliders in hard disk drives. ASME J. Tribol. **124**, 562 (2002)

Accuracy of Surface Topography Characterization Tools

Sirichanok Chanbai, Mark Weber
NanoFocus AG, Oberhausen, Germany

Synonyms

MU, Measurement uncertainty; P, Precision

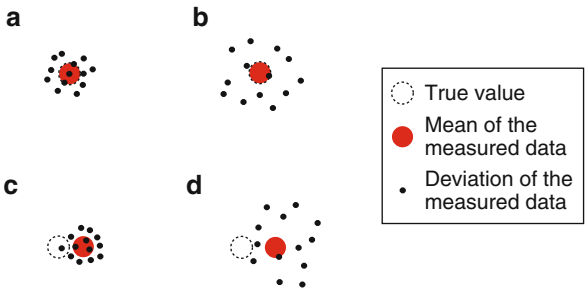
Definition

Accuracy refers to the ability to correctly estimate true value. In surface topography characterization, the accuracy can be interpreted as the capability of an instrument as a whole to predict the true value of a surface topography. The measure of an instrument's accuracy can be literally indicated using the statistical relation of the measured values with respect to the true value of a calibrated standard.

Scientific Fundamentals

Accuracy Versus Precision

In the fields of engineering, industry, and statistics, the *accuracy* of a measurement system is the degree of correctness of the measurement in quantity to its true value. The *precision* of a measurement system, also called repeatability, is the degree to which repeated measurements under unchanged conditions show the same results (Hofer et al. 2005). Although the two words are metaphorical when in use, they are contrary in the context of scientific approach. A measurement system can be accurate but not precise, precise but not accurate, neither, or both. Figure 1 intuitively depicts the accuracy versus precision with respect to a true value. The accuracy is usually affected by the impacts of systematic errors. When a measurement system contains systematic errors, eliminating the systematic errors helps to improve the accuracy but does not enhance the precision. This is due to the fact that the precision strongly relies on all random components and the robustness of an instrument. However, increasing the sampling density practically increases the precision but does not improve the accuracy. Only when a measurement system is both accurate and precise is it reliable. Relevant terminologies such as *measurement uncertainty* and *error* will be discussed in the following section.

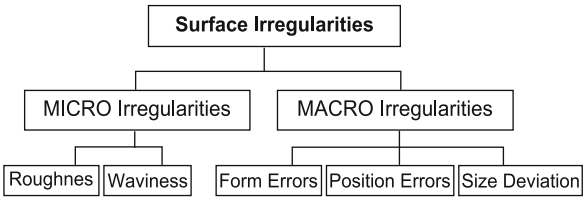


Accuracy of Surface Topography Characterization Tools, Fig. 1 Accuracy versus precision. A measurement system can be: (**a**) accurate and precise, (**b**) accurate but not precise, (**c**) precise but not accurate, (**d**) inaccurate and imprecise

The Measure of Accuracy in Surface Topography Characterization

The measure of "*accuracy*" in surface topography characterization can be quantified by comparing the measured values with the true value of a calibrated artifact. For existing surface topography characterization tools, the calibrated artifacts and methods of calibration have not been thoroughly standardized for a wide span of current instruments. This means that there is still the lack of a common standard in defining the accuracy in each instrument for surface topography characterization. Physikalisch-Technische Bundesanstalt (PTB), for example, provides a standard for calibrated artifacts and methods of calibration for tactile profilometers (TP), but this does not presently apply to optical measurement techniques, i.e., confocal microscope (CM).

One way to quantify the accuracy according to the ISO standard is to introduce accuracy in terms of error or measurement uncertainty. By definition in the ISO standard, accuracy is "*the closeness of the agreement between the*



Accuracy of Surface Topography Characterization Tools, Fig. 2 Surface Irregularities (Udupa et al. 2000)

result of a measurement and the value of the measurand – estimate of the value of specific quantity subject to measurement" (Taylor et al. 1297).

Note that the accuracy of a measurement system for surface topography characterization depends on its resolution and the feature size of the artifacts, in addition to systematic errors in the measurement system and the ambience of measurement. The higher the resolution of a measurement system, the closer the mean of the measured value will be to the true value.

Surface Topography Characteristics

Surface topography are broadly classified into micro and macro scales, as shown in Fig. 2. The surface features, e.g., roughness, waviness, form, lay, laps, tears, and microcracks, greatly affect the performance of a final product. In general, the surface features can be delineated as profiles, areas, forms, and volumes. Moreover, by-products of the surface topography characterization lead to the characterization of the surface functionalities. This provides more information on functionalities of the surface in the forms of friction, contact, lubrication, wear, fatigue strength, tightness of joints, conduction of heat and electrical current, cleanliness, reflectivity, sealing, positional accuracy, load-carrying capacity, resistance to corrosion, and adhesion of paint and coatings (Stout 1994).

Components of Surface Topography

In the spectrum of surface finishes, the long-wavelength component defines the form feature, and the short-wavelength component designates the surface roughness. According to Udupa et al. (2000), three significant components of surface topography resulting from typical machining processes are described as follows.

Form: The general shape of the surface, reflecting the feature variations as a function of roughness and waviness. Normally, the deviations from ideal form are coined as "errors of form" (Udupa et al. 2000).

Waviness: The feature component in which the roughness is superimposed. The waviness component is a result of machine or workpiece deflection, vibrations, chatter, or heat treatment, various causes of strain in the material and extraneous influences (Udupa et al. 2000).

Roughness: The irregularity in the surface that is inherent in the production process, left by the actual machining agent, e.g., a cutting tool, grit or a spark, however, after excluding the waviness and form components (Udupa et al. 2000).

National Metrology Institutes and Standards for Surface Topography Characterization

Deficiency in an international standard for surface topography characterization leads to inconsistency and absence of a common ground for the measure of accuracy in a measurement system. Key factors in surface topography characterization are the calibrated artifacts and the methods of calibration to verify the ability of a measurement system. When there is no common standard for these two significant factors, the accuracy cannot be standardized in validation.

In contact approaches such as tactile profilometers, the ISO and PTB standards for surface topography characterization are acknowledged. However, in non-contact approaches such as optical microscopies, several national standards have been independently developed by individual National Metrology Institutes (NMI), as listed in Table 1.

In tribology, the roughness is the most crucial component among the other components of the surface irregularities. It provides information on wear, lubrication, component life-time prediction, etc. There are standards developed for the evaluation of two-dimensional and three-dimensional surface roughness, which are embraced worldwide. These include ISO 25178, EUR 15178 N, and so on. The standards for the evaluation of three-dimensional roughness parameters are recently developed as an extension from two-dimensional roughness standards.

Uncertainty Defined in ISO Guidelines

Referring to "Guide to the Expression of Uncertainty of Measurements," International Standard Organization (ISO), 1993, the definition of relevant terminologies to accuracy and precision are defined as summarized in Table 2 (Chen et al. 2000). According to the ISO standard, the term *accuracy* should be used with care when quantifying, since this term genuinely represents a qualitative concept. Nevertheless, to quantify the errors of measurements, standard deviation or standard uncertainty is recommended. In this section, the standard uncertainty regulated by ISO is presented. It should be borne in mind

Accuracy of Surface Topography Characterization Tools, Table 1 National Metrology Institutes (NMIs)

NMIs	AIST/NMIJ	CMS/ITRI	DFM	IMGC	KRISS	LNE	METAS
Country	Japan	Taiwan	Denmark	Italy	Korea	France	Switzerland
NMIs	MIKES	NIM	NIST	NMI-VSL	NPL	PTB	
Country	Finland	China	USA	Netherlands	UK	Germany	

Accuracy of Surface Topography Characterization Tools, Table 2 Standard Terminology (Chen et al. 2000)

Accuracy	The closeness of the agreement between the result of a measurement and the value of measurand – estimate of the value of specific quantity subject to measurement (Taylor et al. 1297).
Uncertainty	The estimated possible deviation of the result of measurement from its actual value (Keithley Instruments 1993).
Error	The result of a measurement minus the value of the measurand (Taylor et al. 1297).
Precision	The closeness of agreement between independent test results obtained under stipulated conditions (Taylor et al. 1297).
Repeatability	The closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement (Taylor et al. 1297).
Reproducibility	The closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement (Taylor et al. 1297).
Resolution	A measure of the smallest portion of the signal that can be observed (Keithley Instruments 1993).
Sensitivity	The smallest detectable change in a measurement. The ultimate sensitivity of a measuring instrument depends both on its resolution and the lowest measurement range (Keithley Instruments 1993).

that the uncertainty expresses the quantity of the precision rather than the accuracy.

However, in practice the accuracy is commonly quoted for the measurement errors both in academia and industry usage. This is due to the fact that among the terms *accuracy*, *error* and *uncertainty*, the accuracy and the uncertainty are meaningful to use rather than the error. Several manufacturers specify a range of accuracy by $\pm x$ (unit), but this relatively refers to the measurement error. Some manufacturers have specified the measurement accuracy in terms of 1σ, or 2σ, or 3σ (Chen et al. 2000; Dalton 1998).

Variations, or *measurement uncertainty*, introduced in the measurement data create the differences between the measured and the true value. If these variations are systematic and their causes are known, these variations can be corrected; the accuracy is therefore improved. However, the variations that cannot be corrected are defined as the probability of a "tolerance band" around the true value (Taylor et al. 1297).

- Uncertainty of measurement is *"a parameter associated with the result of measurement, that characterizes the dispersion of values that could be reasonably attributed to the measurand."*

The measurement uncertainty generally consists of several components that in the CIPM (International Committee for Weights and Measures) approach can be grouped with respect to the method used to estimate their numerical values. The measurement uncertainty in the *Guide to the Expression of Uncertainty in Measurement* established by ISO are classified into two types: Type A and Type B, which follows the sources of uncertainty.

A. Those that are evaluated by statistical methods or "component of uncertainty arising from a random effect."
B. Those that are evaluated by other means or "component of uncertainty arising from a systematic effect."

Example: *Evaluating type A standard uncertainties. Uncertainty of x is estimated as the standard deviation of a discrete random variable, or the so-called root-mean-square (RMS) deviation of its values from the arithmetic mean* ($\bar{x}$).

$$u_A(x) = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

where, n is the number of repeated measurements and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This uncertainty is caused by variations of the measured values. Its estimation is not reliable if the number of measurements is low i.e., $n \leq 10$.

General Components of Measurement Uncertainty

1. Standard uncertainties of type A, (υ_A)

The components are found from *statistical distribution of measurement results*, and can be characterized by *experimental standard deviation*; this corresponds basically to random errors in the classical approach. These uncertainties are found by statistical analysis of the results of repeated measurements. Their causes are considered to be unknown and their value decreases with the number of measurements (Taylor and Kuyatt 1994).

2. Standard uncertainties of type B, (υ_A)

Components found from their *expected probability*, i.e., uncertainties of readings of measuring instruments, uncertainties of passive elements, etc. These correspond mainly to the *systematic errors* in the classical approach. These uncertainties are revealed by other means different from the statistical processing of the results of repeated measurements. Therefore, their values do not depend on the number of repeated measurements, but depend on individual source of uncertainty.

3. Combined standard uncertainty of type A and type B, (υ_C)

This uncertainty is calculated from type A standard uncertainty (υ_A) and resulting type B standard uncertainty (υ_B) as:

$$u_C(x) = \sqrt{u_A^2 + u_B^2} \tag{2}$$

4. Expanded uncertainty, ($U(x)$)

The expanded uncertainty $U(x)$ is defined as the product of combined standard uncertainty (υ_C) and the coverage factor (k), expressed by:

$$U(x) = k u_C(x) \tag{3}$$

where,

U is the expanded uncertainty
k is the coverage factor
u_C is the combined standard uncertainty
x is the measured quantity

Note: The coverage factor k most frequently equals 2. In some cases the value of k lie in the interval (Bhushan 1995; Bray et al. 1995). The probability of which the value of the measured quantity lies in the interval defined by expanded uncertainty is 95%. For normal probability distribution, k is set to 2.

In two-dimensional roughness measurements, the root mean square of roughness parameter (R_q) can be used to indicate uncertainty of an instrument.

***Example**: R_q and measurement uncertainty. The deviation from the mean line to the contours of the profile height z at x coordinates discretizing with N sampling points is used to calculate the root mean square roughness parameter (R_q) in 2D, as expressed by:*

$$R_q = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (z_i - \bar{z})^2} \tag{4}$$

When R_q is obtained on a standardized flat-surface artifact, R_q is used to define the uncertainty of the instrument. This uncertainty is dominated by statistical noise if all systematic errors are practically corrected.

Tools for Surface Topography Characterization

Investigating approaches for surface topography characterization can be grouped into two categories: *contact method* and *non-contact method*. In addition, *computational method* is worth noting here as a supplementary tool to characterize surface topography. The computational techniques are applied together with the tools for surface topography characterization to cope with multi-scales of the surface features classification.

Contact Method

In *contact method*, measuring tools are physically contacted the specimen. Tactile profilometry and scanning probe microscopy (SPM) are classified in this category. A tactile profilometer is capable of investigating surface topography in microscopic scale, both in 2D and 3D. Its principle is to scan a tip across the surface with pressing force. Height information of the surface is directly evaluated from the height-variation of the tip (Whitehouse 1997). Thus, roughness profile is revealed by the height value with respect to each lateral sampling point. However, disadvantages of the TP are the limitations related to scanning tip geometries (Conroy and Armstrong 2005). In other words, a scanning tip functions as a filter to cut off

high frequency of the surface features, which provide details of fine surface features. Additionally, the pressing force of the tip causes damages or scan tracks on delicate surfaces.

Nonetheless, TPs are incapable of measuring down to nanometer scale. Scanning probe microscopes, e.g., scanning tunneling microscopes (STM) and atomic force microscopes (AFM), are employed instead to examine fine details in atomic and nanometer scale (Whitehouse 1997). STM functions relatively similar to TP, using a metal tip to scan across the surface, leaving a small gap above the surface, ≈0.2 μm. This is the defined range of the detached distance (Bhushan 1995). When the voltage is applied between two electrodes, the metal tip and an electrically conducted specimen, a tunneling current is induced and flows through the gap. The tunneling current is varied to hold the scanning tip at a given height level over the course of scanning. As a result, the tunneling current variation is a function of the surface topography. In the case when a feedback control system is applied to the STM, the tunneling current can be maintained constant. Then, the allowing movements of the tip are read out as the surface height topography.

Due to the restriction of the STM to electrically conductive specimens, the atomic force microscope supersedes STM in a wide range of the surface types. The AFM are commonly carried out in three modes of operation: *contact mode, non-contact mode*, and *intermittent-contact mode*. In contact mode AFM, a sharp tip, having a radius of approximately 20–50 nm (Bhushan 1995) attached to one end of a low stiffness cantilever, is scanned over the specimen. The contact force is held constant in this mode in the order of 1 nN. The contact force or *normal force* applied on the cantilever generates the deflection of the cantilever in respect to surface topography. In the end, the surface height is calculated using the cantilever's spring constant and the deflection variation (Bray et al. 1995). In non-contact mode AFM, a detached distance of about a nanometer between a sharp tip and the specimen remains constant, i.e., maintained by applied voltage. As a result, the surface height is realized by the deflection variation of the cantilever. However, low-level signal in non-contact mode AFM can be caused by low contact force and high stiffness of the cantilever. To gain advantages from both the contact mode and the non-contact mode AFM, the intermittent contact mode AFM is established. In intermittent contact mode AFM, the cantilever is oscillated. The oscillated cantilever helps reduce the lateral force that is primarily responsible for most of the damage made by the scanning tip (Bray et al. 1995). Then, less destructive and more efficient measurements can be manifested. AFM can also be used to obtain other parameters associated with forces, such as magnetism, electrostatic attraction, chemical attraction, adhesion, friction, wear, and lubrication.

Non-Contact Method

Non-destructive methods are of great interest to examine delicate surfaces. The *non-contact method* based on optical technique can fulfill this requirement. Optical based methods, e.g., confocal microscope and white-light interferometry (WLI) are capable of characterizing surface topography without making any physical contact to the surface. Their use is growing due to speed and the ability to carry out a non-contact measurement over high measurement ranges with high resolutions. Due to laser-speckle effects, white light light-emitting-diode (LED) is an optional light source applied in optical microscopes such as in reflection CMs for roughness measurement. A confocal microscope is basically performed under a focus-detection approach. It differs from a conventional microscope in a way that the observing area is reduced to a single focal point at a time. To render the whole surface, the focal point needs to be scanned (Wilson 1990). In the lateral plane, rotating of the Nipkow disk facilitates generation of a raster scan of multi focal points, with a speed of real-time measurement. To scan in the axial direction, z or height direction, this is achieved by using a piezo-electrical transducer (Jordan et al. 1998). In a CM, a focal point functions as a sharp tip in SPM. The focal radius of the diffraction limit is a function of the numerical aperture of an objective, which defines the lateral resolution optically. Through scanning in the z-direction, a stack of the lateral plane data piled up in the vertical direction contains height information of the surface topography with respect to the lateral sampling coordinates (x, y). Finally, the surface height is evaluated from the distribution of light intensity output, which corresponds to the maximum peak intensity.

For smooth surfaces, white light interferometry is suitable for surface roughness measurements at high resolution. This measuring principle relies on creating interferograms known as *fringes*, the combination of bright and dark bands. The fringes are created by interfering of the light from at least two beams; one scattered from the specimen and the other one from a reference mirror inside the WLI. When the surface is in focus, the fringes contrast is maximum. The interference pattern is formed as a function of the path length difference between the two light beams. For example, if both beams travel exactly the same path length, the beams will interfere with each other

constructively. The number of fringes and their spacing depend on the relative tilt angle between the sample and the reference mirror. Throughout the z scan, a stack of intensity output per frame is recorded by a charge-coupled device (CCD). Thereafter, analyzing the data via interferometric phase mapping programs, height information at each lateral sampling point on the surface is determined (Hariharan 1996).

Computational Method

To provide a more intuitive way of surface functionality manifestation, the *computational method*, has been developed. Basically, tribological surfaces possess multi-scale traits of morphology. This shows in different length scales of the surface features. However, traditional tools to characterize surface topography can carry out only a one-range scale of surface features. This is contrary to the multi-scale nature of tribological surfaces (Stachowiak and Podsiadlo 2004). For example, the surface roughness acquired by a conventional measuring tool results from the filtering of the data in a one-range scale. During the last few decades, computational methods have been extensively studied to cope with the multi-scale morphology characterization. Methodologies to classify surface features computationally include quantitative correlation, neural network, computer-aided systematic particle analysis procedure, scale-invariant pattern recognition, and the hybrid fractal-wavelet method (Stachowiak and Podsiadlo 2004).

Key Applications

Examples of Accuracy Measure Based on non-standardized Methods

Assessment of the Accuracy in a Tactile Profilometer and a While-Light Interferometer Based on Probability

It has been reported by Kohles et al. (2004) that the accuracy of the surface topography characterization tools was assessed by comparing the measured data with the verified true value of the specimen obtained using a SEM. Based on their study, R_a, 2D mean roughness parameter, was used as a basis parameter for comparison. Commercially pure titanium disks were applied as samples. The disks were investigated by a TP and a WLI in diametrical direction. In addition, the true value of surface roughness was verified using a SEM, by measuring cross-sectional profiles. The disk samples were physically cut in diameter, and its cross-sectional surface was polished before investigated by the SEM.

The accuracy of the TP and the WLI was indicated in terms of probability, by which the measured values were compared to the R_a based SEM value. Finally, the precision of the instruments was quantified using the coefficient of variation.

In this case, these probabilities not only indicate the accuracy of individual instrument, but also indicated variables dependency. The variables' dependency includes surface treatments (i.e., machined, acid-etched, titanium plasma-sprayed, and hydroxyapatite-(HA) coated), instruments (TP and WLI), measurement site, and measurement sequence.

Qualitative Measure of the Accuracy in a Surface Reconstruction Algorithm, Obtained by Assessing an AFM's Tip Radius

Influences of the tip radius and sampling interval are of interest and need to be optimized to improve the AFM nanometer-scale measurement efficiency. In an investigation by Chen and Huang (2007), a numerical simulation technique was applied to estimate the tip radius based upon a known rough surface. Due to the difficulties in conducting a real experiment, the optimization was performed using a simulation technique. A 2D measured profile, characterized by Gaussian distribution, was regenerated by a mathematical morphology algorithm in order to function as a surface reference. Thereafter, the relevant tip shape was reconstructed from the reference surface using the blind reconstruction method. The accuracy of the reconstruction algorithm was assessed qualitatively by comparing the simulated tip shape to the real tip. It was deduced that the accuracy in surface reconstruction using the blind reconstruction algorithm can be qualified using the standard deviation and the correlation length of the surface and tip radius.

Accuracy Assessment Using Percentage of Deviation

The accuracy assessment in surface topography characterization using percentage of deviation has been carried out by Ohlsson et al. (2001). The roughness measured using a tactile profilometer and a white-light interferometer were compared to the true values verified by an AFM. The comparison was quantified as a function of the percentage of deviation. To practically compare the results obtained using these instruments, the measurements should be conducted precisely at the same locations on the surface. Therefore, Vickers hardness indentation technique was used to mark the locations. 3D surface roughness parameters, e.g., S_a, S_q, S_z, and S_p, were calculated and used as basis parameters for comparison. The specimens

used in this investigation were plateau-honed cylinder liner, hardened and polished steel roller, honed gear surface, and ECD steel sheet. This assessment method used is similar to the approach executed in the study reported by Kohles et al. (2004).

Surface Topography Characterization in Key Industries

Cold Rolled Sheet Metals

- *characterization of the surface topography, evaluation of 3D surface parameters to describe the metal forming tribological behavior and the pattern of closed-void areas.*

Metal sheet forming processes still encounter great challenges. Its surface topography characteristic leads to a high potential for process and product optimization. Surface topography characterization of cold rolled sheets published by Betalha and Filho (Batalha and Filho 2001), for example, shows excessive studies in this field.

Engine Cylinder Wall

- *characterization of the microstructure surface, the groove angles of the honed surface, and the real volume.*

The microstructure is important in describing the tribological characteristics of the piston and the cylinder wall (Windecker 2001). Windecker has reported the use of white-light interferometer to investigate the surface of cylinder walls. The results proved that optical measurement of roughness values corresponded very well to the tactile measurement (Windecker 2001).

Microelectronics and Semiconductor Devices

- *characterization of the silicon surface roughness, the implant carrier distribution, the defect in layers, and the submicron geometry devices and mask development.*

Surfaces, materials and so forth in microelectronic device are now widely characterized using SPMs such as atomic force, magnetic force, and lateral force microscopes, etc. Scanning probe microscopy is one of the breakthrough technologies that lead to the three-dimensional imaging and the measurement of structures from sub-micrometer to atomic scale. In the study of Natarajan et al. (1997), surface topography of silicon was investigated using AFMs in order to study a few monolayers of silicon oxide, to analyze contaminated IC surface, and to interpret physical phenomena of adhesion between epoxy mold component and BGA solder resists.

Cross-References

- ▶ Atomic Force Microscopy (AFM)
- ▶ Confocal Microscopy
- ▶ Optical Interferometry
- ▶ Scanning Electron Microscopy (SEM)
- ▶ Surface Analysis Using Contact Mode AFM
- ▶ Surface Analysis Using Dynamic AFM
- ▶ Surface Characterization and Description
- ▶ Surface Roughness
- ▶ Surface Variation in Tribological Processes

References

G.F. Batalha, M.S. Filho, Quantitative characterization of the surface topography of cold rolled sheets – new approaches and possibilities. J. Mater. Process. Technol. **113**, 732–738 (2001)

B. Bhushan (ed.), *Handbook of Micro/Nanotribology* (CRC-Press, BocaRaton, FL, 1995)

M.T. Bray, S.H. Cohen, M.L. Lightbody (eds.), *Atomic Force Microscopy/Scanning Tunneling Microscopy* (Springer, New York, 1995)

Yh Chen, Wh Huang, Some issues on atomic force microscopy based surface characterization. Optoelectron Lett **3**, 129–132 (2007)

F. Chen, G.M. Brown, M. Song, Overview of three-dimensional shape measurement using optical methods. Opt Eng **39**, 10–22 (2000)

M. Conroy, J. Armstrong, A comparison of surface metrology techniques. J. Phys.: Confer Ser **13**, 458–465 (2005)

G. Dalton, Reverse engineering using laser metrology. Sensor. Rev. **18**, 92–96 (1998)

P. Hariharan, *Handbook of Optics, Interferometers, Chapter 21* (McGraw-Hill, New York, 1996)

M. Hofer, G. Strauß, K. Koulechov, A. Dietz, Computer assisted radiology and surgery – evaluating cas-systems, in *CARS 2005: Computer Assisted Radiology and Surgery*, volume 1281 of *International Congress Series*, (Computer Assisted Radiology and Surgery and Elsevier, Amsterdam, May 2005) pp 548–552

H.-J. Jordan, M. Wegner, H. Tiziani, Highly accurate non-contact characterization of engineering surfaces using confocal microscopy. Meas. Sci. Technol **9**, 1142–1151 (1998)

Keithley Instruments, *Low Level Measurements* (Keithley Instruments, Inc., Cleveland, OH, 1993)

S.S. Kohles, M.B. Clark, C.A. Brown, J.N. Kenealy, Direct assessment of profilometric roughness variability from typical implant surface types. Int. J. Oral Maxillofac. Implants **19**, 510–516 (2004)

A. Natarajan, C.Q. Cui, D.P. Poener, M.K. Radhakrishnan, Applications of atomic force microscopy for semiconductor device and package characterization. in *Physical and Failure Analysis of Integrated Circuits, Proceedings of the 1997 6th International Symposium*, (IPFA '97, Singapore, July 1997) pp 275–279

R. Ohlsson, A. Wihlborg, H. Westberg, The accuracy of fast 3d topography measurements. Int. J. Mach. Tool. Manufact. **41**, 1899–1907 (2001)

G.W. Stachowiak, P. Podsiadlo, Classification of tribological surfaces. Tribol. Int. **37**, 211–217 (2004)

K.J. Stout, *Three Dimensional Surface Topography: Measurement Interpretation and Applications* (Penton Press, London, 1994)

B. N. Taylor, C. E. Kuyatt, Guidelines for evaluating and expressing the uncertainty of NIST measurement results. in *Technical report, NIST Technical Note 1297*, (U.S. Government Printing Office, Washington, DC, 1994)

G. Udupa, M. Singaperumal, R.S. Sirohi, M.P. Kothiyal, Characterization of surface topography by confocal microscopy: I. principles and the measurement system. Meas. Sci. Technol. **11**, 305–314 (2000)
D.J. Whitehouse, Surface metrology. Meas. Sci. Technol. **8**, 956–972 (1997)
T. Wilson, *Confocal Microscopy* (Academic Press, London, 1990)
R. Windecker, High resolution optical sensor for the inspection of engine cylinder walls. Optik **112**, 407–412 (2001)

Acid Phosphates

► Ashless Phosphate Esters

Acoustic Emission

► Air Bearing Diagnosis

Active Lift Seal

► Mechanical Seals

Active Tribology

► Tribotronics – Monitoring-Based Active Friction Control

Adaptive Hard Coatings Design Based on the Concept of Self-Organization During Friction

G. Fox-Rabinovich
Department of Mechanical Engineering, McMaster University, Hamilton, ON, Canada

Synonyms

Hard coatings with adaptive behavior

Definition

Adaptive hard coatings are and emerging generation of wear-resistant coatings. Adaptability of any material is related to the changes in its characteristics in response to an external stimulus, which leads to weakening of the effects of its influence. Application of adaptive coatings enables a tribo-system to shift to a milder wear mode due to formation of protective/lubricious tribo-films as a result of interaction with the external environment under operation.

Scientific Fundamentals

Friction is a complex phenomenon that is associated with a variety of different mechanical, physical, and chemical processes (Hardwicke 2003). Traditionally, this area of applied science and engineering has been examined mainly via a purely mechanical approach. However, a few decades ago, friction began to be looked at with a more generic concept of complexity (Beckerman 2000).

One achievement of modern physics is the development of concepts that deal with the complexity of nature. Major progress in this area is associated with the Nobel Prize winner I. Prigogine (1980). Prigogine's ideas were further developed by H. Haken, who introduced the term "synergistics" to the field of applied physics (Haken 1987). Synergistics is a scientific discipline that considers the processes related to the self-organization phenomena, stability and degradation of the structures that form under conditions that are far from equilibrium. It deals with the spontaneous formation of spatial, temporal, and functional structures (Mansson and Lindgren 1990). The irreversible processes of these structures' formation are associated with non-equilibrium phase transformations. These transformations are associated with specific bifurcation or instability points where the macroscopic behavior of the system changes qualitatively and may leap either into chaos or into greater complexity and stability (Beckerman 2000; Fox-Rabinovich and Totten 2006). In the latter case, as soon as the system passes these specific points, its properties change spontaneously due to self-organization and formation of dissipative structures. The driving force of the self-organization process is that the open systems aim to decrease production of entropy during nonstationary processes. Spontaneous formation of the dissipative structures as a result of symmetry disruption can be realized only in open systems, which exchange energy, matter, and entropy with their environment (Fox-Rabinovich and Totten 2006). This phenomenon is the focus of attention for researchers in many different fields of science, including tribology.

All mechanical, chemical, and physical processes that are developing during friction follow the generic laws of irreversible thermodynamics. Tribology has made significant progress in understanding the complex phenomena

associated with friction and wear based on the ideas of irreversible thermodynamics and self-organization phenomena. Self-organization usually requires severe frictional conditions. The theory of self-organization, in association with severe tribological conditions, has had impressive confirmation that has lead to valuable practical results (Kostetskaya 1985). Thus, heavy loaded tribo-systems (HLTS), working under severe conditions associated with high temperature and stresses and having an intensive wear rate, are considered in this essay. Cutting and stamping tools are excellent examples of HLTS. Efficient protection of the friction surface is critical in the severe conditions associated with modern cutting and stamping operations. Hard plasma vapor deposited (PVD) coatings are best suited for this application.

The principles of synergistics can be successfully applied in the development of advanced tribo-systems and materials. These principles provide researchers with powerful methodological tools to enhance beneficial non-equilibrium processes. The tribo-processes under consideration result in the ability of the friction surface to exhibit self-protection and self-healing properties. Assuming these principles, scientists have an opportunity to develop a new generation of tribo-systems or/and materials that exhibit adaptive, or smart, behavior. An adaptive/smart tribo-system is the one that responds to the external mechanical, thermal, and chemical forces with a positive feedback loop that leads to an improvement in the wear characteristics of a tribo-couple (Fox-Rabinovich and Totten 2006).

As outlined above, interaction between frictional bodies was traditionally thought to lie wholly in the domain of the mechanical response of a system to these conditions; relatively little attention was paid to the role of physical and chemical interactions between frictional bodies. Recent research has challenged the latter viewpoint, and it is now realized that extensive physical as well as chemical interactions can occur, both between frictional bodies as well as with the surrounding atmosphere. These considerations, combined with approaches of irreversible thermodynamics and self-organization, resulted in critical progress in tribology (Fox-Rabinovich and Totten 2006).

This essay explains how these generic ideas work for HLTS and discusses the application of synergistics to tribology and the role played by physico-chemical interactions in modifying and controlling friction and wear during operation. This is of even greater importance due to the demand for higher productivity, often under conditions where a lubricant cannot be used (such as in dry machining conditions).

The increasing demands of modern engineering have spawned the development of new adaptive materials for use as HLTS, possessing a high level of wear resistance. The development of adaptive surface-engineered materials is a typical problem of engineering optimization. In this process, an integrated engineering and physical approach to the problem of developing of novel wear-resistant materials must be taken. The key concept is associated with the tribological compatibility of two surfaces interacting at friction.

A concept of tribological compatibility, introduced by N. A. Bushe (Fox-Rabinovich and Totten 2006), can be defined as follows. Tribological compatibility is the ability of a tribo-system to provide optimum friction conditions within the given range of operating conditions using the chosen criteria (Fox-Rabinovich and Totten 2006). Tribological compatibility is related to the capacity of the two surfaces to adapt to each other during friction, providing wear stability without surface damage to the two components of the specific tribo-system for the longest (or a given) period of time. The goal is to achieve stable tool service and a predictable rate of wear with a given set of operating parameters. In this interpretation, as outlined above, compatibility implies an integrated optimization, both from an engineering (minimal wear rate) and physical (self-organizing) point of view.

Modern tribology, an interdisciplinary science based on mechanics, physics, chemistry, materials science, metallurgy, etc., is a very complex subject. Models that generalize our knowledge in this area of science and are acceptable for engineering applications are needed. Because friction is a process of transformation and dissipation of mechanical energy into other kinds of energy, from our point of view, an energy-based approach is the most effective.

A tribo-system can be considered to be an open thermodynamic system. For these systems, the second law of thermodynamics is still operative, but a more complex and general behavior is found compared with the more classical case discussed in standard textbooks. According to the principles developed by I. Prigogine (1980), the second law does not eliminate the possibility of highly organized dissipative structures being formed in an open tribo-system.

During friction and wear, adaptation of the materials of the tribo-couple takes place and, in many cases, leads to drastic structural changes within the surface layers. The characteristics of surface and under-surface layers (such as geometrical parameters, microstructure, and physico-chemical and mechanical properties) change during the process of adaptation as a response to external impact.

The adaptation is completed in the initial stage of the life of a tribo-system, i.e., during the running-in stage. Although evolving in a step-by-step fashion and becoming increasingly complicated during this stage, the tribo-films (secondary structures) eventually stabilize for a given tribo-pair and conditions of friction. When the characteristics of the surface layers become optimal, the running-in phase is completed, and the parameters of friction (i.e., the coefficient of friction and the wear rate) stabilize at a lower level (Fox-Rabinovich and Totten 2006). The process of wear transforms to a post running-in, steady stage.

During the steady stage of wear, the tribological compatibility is controlled by the stability of the tribo-films formed and wear resistance of the contacting materials, their fatigue life as well as lubricity of the surface layers. The tribological compatibility, especially for HLTS, depends on the intensity of surface damage during the initial (running-in) stage. Special coatings to prevent severe surface damage during initial stages of wear are used.

The self-organizing phenomenon (SO) is characterized by the formation of thin (from several nanometers up to a micron thick) tribo-films at the friction surface. These are generated from the base material by structural modification and/or by interaction with the environment. It has been estimated that a significant portion of the work of friction can be accumulated in the tribo-films. Thus, tribo-films represent an energy sink for the preferential dissipation of the work of friction (Fox-Rabinovich and Totten 2006).

The synergistic processes of adaptation to the extreme deformation, thermal, and diffusive conditions associated with friction can be concentrated in the thin layer of tribo-films. Self-organizing of the tribosystem is often accompanied by a kinetic phase transition. In this case, all the interactions are localized in a thin surface layer, the depth of which can be lower by an order of magnitude than that typically associated with damage phenomena. The rate of diffusion and chemical reactions may also increase substantially, while the surface layers may become ductile. In addition, the solubility of many elements might be increased and non-stoichiometric compounds might form.

There are two kinds of tribo-films: (1) super-ductile and lubricious and (2) tribo-ceramics with thermal barrier properties and, generally, increased strength (Fox-Rabinovich and Totten 2006). Tribo-films of the first type are observed after structural activation that is marked by an increase in the density of atomic defects at the surface. The tribo-films are supersaturated solid solutions formed by reaction with elements from the environment (most often, oxygen from the environment). The reaction between oxygen and the substrate during the self-organizing process is very different from the classical case encountered in regular oxidation experiments. Tribo-films of the first kind are similar to Beilby layers, having a fragmented and textured structure, aligned in the shear direction (Fox-Rabinovich and Totten 2006). In these secondary structures, the material may be super-plastic (due to a very fine-grained or amorphous-like structure), with an elongation up to 2,000%. The amorphous-like structure of the tribo-films may also lead to a decrease in the heat conductivity of the surface, an important consideration in controlling friction of cutting tools (Fox-Rabinovich and Totten 2006). These tribo-films promote energy dissipation during friction. In the beginning, this energy dissipation can result in an entropy production increase within a zone of friction. But these tribo-films' lubricious-like action leads to prevention of subsurface layer damage and eventually results in an entropy production decrease within the zone of friction.

The tribo-films of the second type (tribo-ceramics containing a higher percentage of elements such as oxygen) are mainly formed by thermally activated processes. These tribo-films are usually non-stoichiometric compounds (Fox-Rabinovich and Totten 2006). Tribo-films of this type have high thermodynamic stability, thermal barrier properties, and, generally, a high hardness. It is thought that one of the benefits they bestow is to accommodate the stress associated with friction by elastic rather than plastic deformation. The adaptation of the tribo-system in this case relies upon a low-intensity interaction with a frictional body in contact and the high hardness of the thin surface film formed during cutting. This results in low entropy production during friction, which leads to a decrease in wear rate. On the other hand, destruction of these hard films should be prevented due to proper engineering of the substrate material that ensures effective support of the tribo-films during friction. External impact during friction inevitably results in the tribo-films' destruction, however, the same impacts as well as mass transfer with their environments leads to their permanent regeneration. Stability of the self-organization process is caused by the dynamic self-regulation of the process of tribo-films' generation and destruction.

The entire diversity of processes that take place during friction can be divided into two groups: quasi-equilibrium, steady-state processes (encountered during post running-in normal friction and wear) and non-equilibrium, unsteady state, associated with surface damage processes. Surface damage is usually observed in the

initial (running-in) and final (avalanche-like) stages of wear. During the period of service under stable (normal) friction and wear conditions, no macroscopic damage of the contact surfaces can be observed. Stable friction corresponds here to the completion of the self-organizing processes discussed above and the transition to a post running-in stage of wear.

The major goals of the friction control are (1) to decrease wear and surface damage intensity within the running-in stage of the wear process and (2) to widen the stable wear stage. This can be done by means of surface engineering (Kostetskaya 1985).

Modern technologies of surface engineering, such as PVD coatings, give the initial surface a high density of lattice imperfections (Fox-Rabinovich and Totten 2006). Thus, a surface has a highly non-equilibrium state that dramatically enhances beneficial processes of mass transfer and formation of protective tribo-films (Knotek et al. 1994; Kostetskaya 1985). The only limitation is the ability of this highly activated surface to accumulate the additional structure imperfections that are generated during friction without deep surface damage. The initially highly activated non-equilibrium surface, such as nanocrystalline coatings (Fox-Rabinovich and Totten 2006), leads to explosion-like mass transfer followed by formation of protective tribo-films. Once the self-organization process is completed, the friction characteristics change critically and the wear process transforms to the after running-in, stable stage.

Friction control in this context implies the existence of a stable tribo-system, which resists any instability, leading to the deep surface damage (Fox-Rabinovich and Totten 2006). The transition from a thermodynamically non-equilibrium condition to a more stable, quasi-equilibrium condition is connected to the accelerated formation of a beneficial surface structure formed as a result of self-organizing. Any tribo-system combines some features of an artificial system (engineering system, mechanism) and a natural phenomenon (the friction by itself). From the point of view of self-organizing, both natural and synthetic processes can be considered during friction. Therefore, it is necessary to control (or modify) the synthetic processes to encourage the evolution of those natural processes that lead to a minimum wear rate. The problem of tribological compatibility includes developments that ensure the stabilization of the friction and wear parameters, in particular, by the development of adaptive surface engineered material.

There are two concepts of surface engineering. One is to create a coating with a limited number of advanced properties that gives the tribo-system that ability to sustain severe external impact with no or minimal change in the coatings properties (Veprek and Argon 2001). The other approach is based on the adaptability of tribo-systems. Adaptive coating application weakens internal impact and allows shifting of the process to a milder wear mode. Based on long-term research, it was concluded that this is the better method of friction control under severe frictional conditions such as, for example, high-performance machining (Fox-Rabinovich and Totten 2006; Fox-Rabinovich et al. 2008). Eventually, tribological compatibility during cutting leads to two major practical improvements: (1) tool life enhancement and (2) improved workpiece quality (e.g., a better surface finish, improved dimensional accuracy).

Key Applications

A major topic of this essay is surface engineered materials, the reason being that state-of-the-art surface engineering provides exceptional opportunities for the development of novel materials. State-of-the-art nano-crystalline/composite/laminated PVD coatings usually have a highly non-equilibrium state that gives them some outstanding properties, which usually cannon be realized using traditional methods of the materials fabrication. These novel surface-engineered materials are perfectly suited to their severe applications. Modern extreme applications such as high performance machining also imply far away from equilibrium high temperatures/stresses conditions where excessive wear occurs most often. To meet these requirements, a new generation of materials that can resist this external impact is needed. These include advanced surface-engineered materials with nano-scale structure. Application of adaptive surface-engineered materials critically enhances the resistance of the friction surface against severe external attack that traditional materials cannot sustain.

Another novel direction in material science relates to the final point of friction control, which is ensuring conditions where elements of the tribo-system can work in synergy. This can be accomplished due to the development of synergistically alloyed surface engineered materials.

Modern tribological applications tend to be increasingly severe and often extreme. A number of processes occur simultaneously, resulting in complex frictional conditions (Fox-Rabinovich and Totten 2006). There are a large number of simultaneously occurring wear mechanisms on the surface of a cutting tool under high performance machining conditions (high cutting speed). For example, for conditions with interrupted cutting (Knotek et al. 1994), in addition to intensive adhesion, abrasion,

and tribochemical reactions at high temperatures there is fatigue impact combined with thermal shock. To address these issues, novel surface engineering methods must be multi-functional.

Multi-functional wear-resistant coatings and films are complex systems. System science employs two distinct approaches to complex system engineering (Beckerman 2000). One is the traditional reductionist approach and the second in a concept of emergence. It must be emphasized that the reductionist approach has been quite beneficial thus far in the field of surface engineering. As is outlined above, one or two properties are usually emphasized for hard coatings, such as hardness and oxidation stability (Veprek and Argon 2001), and major progress has been made thus far in achieving these particular properties (Veprek and Argon 2001). Other properties are largely ignored, however, and, for this reason, progress in harder coating development does not guarantee multi-functionality (Fox-Rabinovich and Totten 2006). The concept of emergence is intrinsic to all types of systems (Beckerman 2000). It has been embraced by the soft sciences (biology, sociology, etc.) but has been largely ignored by the engineering community, under the mistaken belief that engineered systems do not display it (Beckerman 2000). Of course, this concept can be used in a realistic way, but this concept might represent a new strategy in multi-functional coatings design. It is worth noting that nobody can ignore properties of hard coatings outlined above. However, based on the complexity of the cutting tool wear process, advances—even significant advances—in a limited number of characteristics (usually at the cost of the other properties) do not automatically result in superior wear resistance under various operating conditions (Fox-Rabinovich and Totten 2006). Hardness of definite level is important, no doubt, but this should not be the entire goal of optimization, especially at a cost to the other properties. An integrative, holistic approach (Beckerman 2000) can be a more beneficial way to achieve multi-functionality of the coating, which is related to its emergent properties (Fox-Rabinovich and Totten 2006; Fox-Rabinovich et al. 2008).

As outlined above, to reduce entropy production during friction, a surface-engineered layer should posses a highly non-equilibrium state (Fox-Rabinovich and Totten 2006), which is typical of a nano-structured coating. Therefore, complexity of the surface-engineered system is related to the non-equilibrium state that gives this system the ability to better adapt to strongly varying operating conditions.

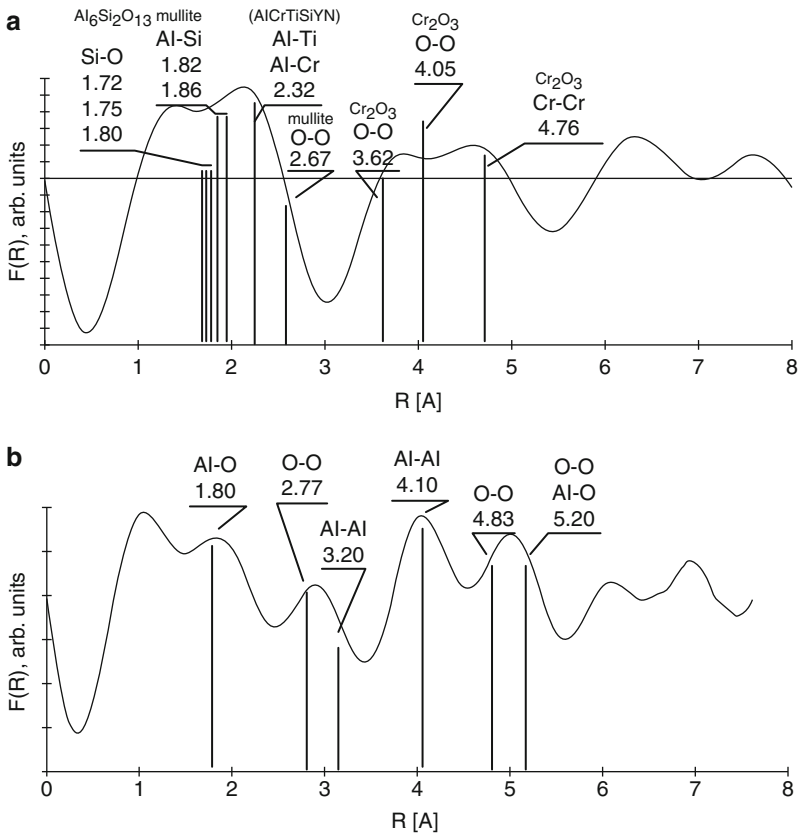
An issue in modern manufacturing is the high-performance machining of hard-to-cut materials. Major features of the typical operating conditions are high temperature/stress on the friction surface complicated by the adhesive interaction with the work-piece and harsh oxidative attack from the environment. Typical examples of modern machining of hard-to-cut materials are high-speed cutting of hardened tool steels and advanced aerospace nickel-based superalloys.

Adaptation during friction leads to conditions when further service of the material takes place with a stable minimal entropy production with the formation of newly generated layers of coating (protective and lubricious tribo-films) as a result of interaction with the environment. Under severe operating conditions, interaction with the environment (tribo-oxidation) plays a decisive role in friction control. One way to enhance adaptive characteristics is nano-structuration of the coatings. Nano-structured PVD coating application does not automatically result in the improvement of volumetric properties, such as hardness (Fox-Rabinovich and Totten 2006). On the other hand, the nano-structuration of the coatings strongly enhances their physicochemical properties (Fox-Rabinovich and Totten 2006).

One drawback of hard coatings is a lack of multi-functionality (Fox-Rabinovich and Totten 2006). To achieve this generic characteristic, a coating must obtain a number of features that can be considered as emergent properties or, in other words, the features of the coating work as a whole, in synergy (Fox-Rabinovich et al. 2008). One way to improve wear performance of an adaptive hard coating is to enhance its emergent properties.

The Al-rich TiAlN and AlTiCrN families of nano-crystalline hard PVD coatings are mostly used for high-performance machining applications. Adaptability of the hard coatings has been intensively investigated (Fox-Rabinovich and Totten 2006). It has been shown that adaptive hard PVD coatings mostly outperform other categories of coatings under aggressive and extreme cutting conditions (Fox-Rabinovich and Totten 2006; Fox-Rabinovich et al. 2008). However, it remains a challenge to further optimize the design of adaptive hard coatings for varying and intensifying operating conditions.

A nano-crystalline TiAlCrSiYN coating has been developed that presents some emergent properties under extreme operating conditions (Fox-Rabinovich et al. 2008). The synergistically alloyed TiAlCrSiYN coating has the ability to sustain extreme tribological conditions associated with high-speed machining of hardened tool steels (Fox-Rabinovich et al. 2008). It was shown in our research that the introduction of



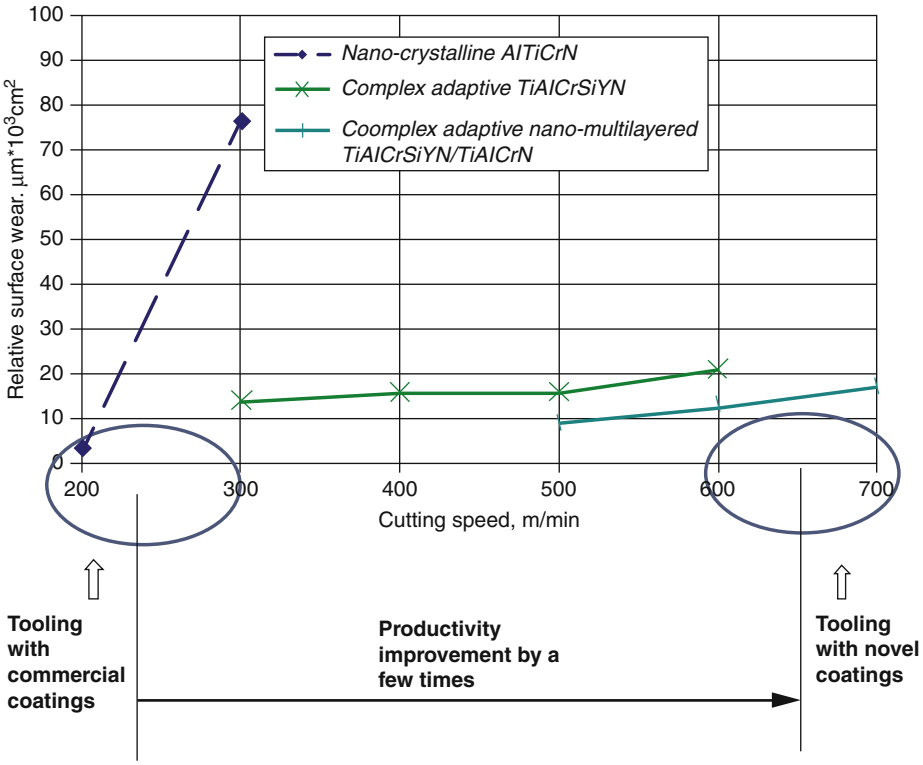
Adaptive Hard Coatings Design Based on the Concept of Self-Organization During Friction, Fig. 1 Fourier transforms of electron energy loss fine spectra obtained from two different places (**a** and **b**) on the worn surface of a cutting tool with TiAlCrSiYN/TiAlCrN nano-multilayered coating. Each peak position corresponds to the radius of coordination sphere in the crystal lattice. Complex chemical composition of the tribo-films provides a large set of pair inter-atomic distances. Two protective phases, namely sapphire- and mullite-like, as well as lubricating Si-O and Cr-O tribo-films co-exist on the friction surface

a more complex design of the nano-multilayered TiAlCrSiYN/TiAlCrN coating could be a way to improve coatings characteristics and enhance their emergent properties.

In more complex systems with a higher amount of interacting processes, self-organization could start under milder frictional conditions. Ideally, it starts with the beginning of friction. Therefore, wear rate growth under increasingly severe operating conditions could be inhibited before the self-organization has begun, and thus an increase in wear of the tribo-system could be prevented.

Increases in the complexity of an engineering system must be controlled based on the concept of the emergent properties of the system. If characteristics of the system are tuned in a way to enhance the emergent properties, then the goal of the entire control is achieved and multifunctionality of the system is improved.

One fascinating direction of modern research is the control of characteristics of tribo-films formed during friction (Fox-Rabinovich and Totten 2006; Fox-Rabinovich et al. 2008). It was found that under extreme operating conditions (a combination of heavy loads and temperatures above 1,000°C), if the coating possesses a strongly non-equilibrium state, then the tribo-films formed have outstanding protective properties that strongly enhance adaptability of hard coatings (Hovsepian et al. 2006). Protective tribo-films with sapphire and mullite structure can form on the friction surface (Fig. 1). The sapphires are known as the hardest of oxide crystals that possess high strength and excellent thermal shock resistance at high temperatures (Fox-Rabinovich et al. 2008). Mullite is a refractory compound that has similar high temperature properties (Fox-Rabinovich et al. 2008). Both materials, especially sapphire, have an excellent ability to



Adaptive Hard Coatings Design Based on the Concept of Self-Organization During Friction, Fig. 2 Relative surface wear (RSW) values versus cutting speed during machining of hardened H 13 tool steel for commercial TiAlCrN and novel adaptive mono-layered TiAlCrSiYN and TiAlCrSiYN/TiAlCrN nano-multilayered coatings

accumulate energy of internal impact (Fox-Rabinovich et al. 2008) that leads to reduction in entropy production during friction. Moreover, in more complex surface-engineered systems such as multi-layered TiAlCrSiYN/TiAlCrN coating, simultaneous formation of a wide variety of the refractory compound tribo-films such as protective sapphire and mullite, as well as lubricating silicon and chromium oxides, takes place (Fig. 1), which leads to excellent protection as well as sufficient lubrication of the surface under extreme frictional conditions (heavy loads combined with high operating temperature, 1,000°C). Chromium and silicon oxides are characterized by good lubricating properties at high temperatures (Fox-Rabinovich and Totten 2006). At the same time, being chemically stable phases, they prevent intensive adhesive interaction at the workpiece/tool interface and, in this way probably reduce heat generation during cutting (Fox-Rabinovich and Totten 2006). Simultaneously protective and lubricious tribo-film formation is probably the best manner of adaptation of the tribo-system to severe frictional conditions.

Long-term studies of the tool life of different categories of hard coatings (nano-composite, nano-crystalline, and mono- and multilayered) show that the adaptive nano-structured TiAlCrN family of coatings has a greater potential for dry, high-performance machining applications (Fox-Rabinovich and Totten 2006; Fox-Rabinovich et al. 2008).

Figure 2 presents values of the relative surface wear (RSW) parameter versus cutting speed. The RSW parameter is calculated as the ratio of the radial dimensional wear of the cutting tool to the area of the machined surface (Fox-Rabinovich et al. 2008). The RSW parameter was used as a generalizing measure of the wear rate of different coatings in relation to intensifying frictional conditions, which occurs in our studies due to a cutting speed increase within a range of 200–700 m/min. The data presented show the comparable efficiency of the different categories of hard PVD coatings from the point of view of the machining process's productivity. Higher values of RSW relate to a lower productivity and vice versa. At the lower range of cutting speed of 200–300 m/min, which is still the

most widely used in industry, the RSW values of the "basic commercial" TiAlCrN coating are similar but slightly better, compared with the other categories of state-of-the-art coatings (Fox-Rabinovich et al. 2008).

However, with a cutting speed growth of up to 500 m/min, the RSW parameter values increase, sometimes critically, for all the commercial coatings studied (Fox-Rabinovich et al. 2008), excluding the TiAlCrSiYN-based coatings. A multi-layered TiAlCrSiYN/TiAlCrN coating shows constant RSW values versus cutting speed up to speeds of 500 and 600 m/min. This means that the productivity of the machining process for the cutting tools with the multilayered TiAlCrSiYN/TiAlCrN coating significantly grows because the machining time drops with the cutting speed while RSW values remain constant. The most exciting feature of this adaptive coating is its ability to work efficiently under increasingly severe operating conditions, which other coatings cannot sustain. To determine limits of the operating conditions for this coating, an intensive speeding up was performed. It was shown that the coating can sustain cutting speeds as high as 700 m/min when the frictional conditions are extreme. Operating parameters that are used in industry and are directly related to the productivity of the machining process, in particular, cutting speed, are a few times lower. Even cutting data presented in the research are significantly lower (Fox-Rabinovich et al. 2008; Hovsepian et al. 2006).

In our opinion, future generations of hard coatings for extreme tribological applications must be complex engineering adaptive systems that posses emergent properties. In this way, they are able to sustain a number of external impacts and adapt to changing (and severe) operating conditions. Coatings with emergent properties have the ability to perform as higher-ordered systems and therefore can achieve previously unattainable life under extreme tribological conditions.

Cross-References

- ► Anti-adhesion/Stiction Surface Design, Fabrication, and Applications
- ► Cutting Tool Wear and Failure Mechanisms
- ► Design of Wear-Resistant Coatings for Engine Components
- ► Manufacturing Tribology
- ► Nanocomposite Coatings
- ► PVD and CVD Coatings
- ► PVD: Cathodic Arc and High Power Impulse Magnetron Sputtering (HIPIMS)
- ► PVD: Ion Plating
- ► Surface Nanocrystallization and Hardening (SNH)

References

L.P. Beckerman, Application of complex systems science to systems engineering. Syst. Eng. **3**(2), 96–102 (2000)

G.S. Fox-Rabinovich, G. Totten, Self-organization during friction: *Advanced Surface Engineered Materials and Systems Design*, ed. by G.S. Fox-Rabinovich, G. Totten (CRC Press/Taylor & Francis, Boca Raton, 2006)

G.S. Fox-Rabinovich, S.C. Veldhuis, G.K. Dosbaeva, K. Yamamoto, I.S. Gershman, A. Kovalev, B.D. Beake, L.S. Shuster, Nano-crystalline coating design for extreme applications based on the concept of complex adaptive behavior. J. Appl. Phys. **103**, 083510 (2008)

H. Haken, Synergetics: an approach to self-organization, in *Self-Organizing Systems*, ed. by F.E. Yates (Plenum Press, New York, 1987)

C.U. Hardwicke, Recent developments in applying smart structural materials. JOM **12**, 15–16 (2003)

P.E. Hovsepian, C. Reinhard, A.P. Ehiasarian, CrAlYN/CrN superlattice coatings deposited by the combined high power impulse magnetron sputtering/unbalanced magnetron sputtering technique. Surf. Coat. Technol. **201**(7), 4105–4110 (2006)

O. Knotek, E. Lugscheider, F. Löffler, G. Krämer, H. Zimmermann, Abrasive wear resistance and cutting performance of complex PVD coatings. Surf. Coat. Technol. **68–69**, 489–493 (1994)

N.B. Kostetskaya, *Structure and energetic criteria of materials and mechanisms wear-resistance evaluation*. Ph.D. dissertation, Kiev State University, Kiev, 1985

B.A. Mansson, K. Lindgren, Thermodynamics, information and structure, in *Nonequilibrium Theory and Extremum Principles*, ed. by S. Sieniutycz, P. Salamon (Taylor & Francis, New York, 1990), pp. 95–98

I. Pigogine, *From Being to Becoming* (W.H. Freeman, San Francisco, 1980)

S. Veprek, A.S. Argon, Mechanical properties of superhard nanocomposites. Surf. Coat. Technol. **146–147**, 175–182 (2001)

Additive Chemistry Testing Methods

Mark T. Devlin
Afton Chemical Corporation, Richmond, VA, USA

Synonyms

Bench Tests; Engine Tests; Rig tests for additives

Definition

Additive and lubricant tests are designed to measure (1) wear and surface damage prevention, (2) deposit control, (3) sludge control, (4) friction control, (5) lubricant degradation, (6) contaminant effects on lubricant performance, and (7) lubricant rheology. These tests can be simple laboratory tests or complex rig/engine and fleet tests. The cost and time involved in these different tests requires that prudent selection of the lubricants being tested be considered when developing additives and lubricants.

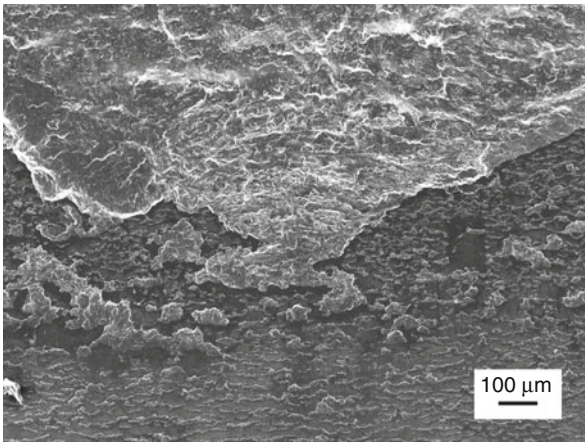
Scientific Fundamentals

The testing of lubricant additive chemistry is closely related to the testing of fully formulated lubricants. While additives can be tested by themselves, most additives only improve one or two aspects of the performance properties required of a fully formulated lubricant. For example, an additive designed to prevent wear may have no effect on the ability of the lubricant to be transported through the mechanical system (lubricant rheology). In addition, interactions between additives and interactions between additives and the base oils in which they are blended often determine the performance of lubricants. For example, an additive designed to prevent wear may interact on a surface with an additive that controls friction. The wear prevention and friction control properties of the two additives together are usually quite different than would be predicted by the performance of each additive separately. Therefore, to best understand additive chemistry testing methods, the general performance features required of a fully formulated lubricant should be understood. In addition, since additive-additive and additive-base oil interactions often control overall lubricant performance, an understanding of how to design lubricant formulations to examine these interactions should be understood (Rudnick 2008).

Lubricants are used to protect mechanical systems and to ensure the efficient operation of the mechanical system. To protect the mechanical system, lubricants must prevent wear and surface damage from fatigue and control deposit (see Fig. 1) and sludge formation. To improve mechanical efficiency, lubricants must control friction. Lubricants must also maintain their ability to protect mechanical systems and control friction as the lubricants are either thermally or oxidatively degraded or when the lubricants are contaminated by water, combustion by-products (in the case of vehicle operation), and air. Contamination of lubricants is a complicated issue since a wide variety of performance properties of the lubricant may be compromised. For example, water may cause additive systems to precipitate out of the lubricant or cause corrosion of metal surfaces. Combustion by-products may cause accelerated degradation of the lubricant or the formation of deposits. Air may cause an increase in the rate of oxidative degradation of the lubricant or cause the collapse of lubricant films that prevent wear by disrupting the flow of lubricant into contact zones. This is clearly not an exhaustive list of contaminant effects on lubricant performance but does give an idea of the complexity of investigating contamination on lubricant performance. Finally, and perhaps most importantly, lubricants must be able to be transported throughout a mechanical system under a wide variety of operating conditions, so the viscosity of lubricants under different temperatures, loads, and shear conditions should be determined (Haycock et al. 2004).

Additive Chemistry Testing Methods, Fig. 1 Deposits formed on pistons during ASTM D7320 engine test

In order to determine lubricant performance, there are tests that can easily be run in a chemical laboratory (bench tests), tests in mechanical systems operating under controlled conditions (rig or engine tests), and tests in mechanical systems operating under actual conditions (field or vehicle tests). There are also more sophisticated analytical techniques that are used to characterize the chemical structure of additives and base oils (for example, infrared spectroscopy and nuclear magnetic resonance spectroscopy) and to characterize films formed by additives on the mating surfaces of mechanical systems (for example, x-ray photoelectron spectroscopy and scanning electron microscopy). These techniques are used to better understand a lubricant's ability to protect mechanical systems and improve efficiency but are not necessary to determine if a lubricant is suitable for use. For example, Fig. 2 shows a scanning electron micrograph of a pit formed on the surface of a spur gear. The pit appears to originate from the band of micropits near the root of the gear tooth. This image suggests that to prevent pit formation on gear teeth, the formation of micropits must be minimized. This type of analysis helps lubricant developers better understand the mechanism by which surface damage occurs. However, in this case, the performance of the lubricant is determined by the ability to prevent the formation of surface fatigue damage. A better understanding of the mechanism by which surface pits form allows



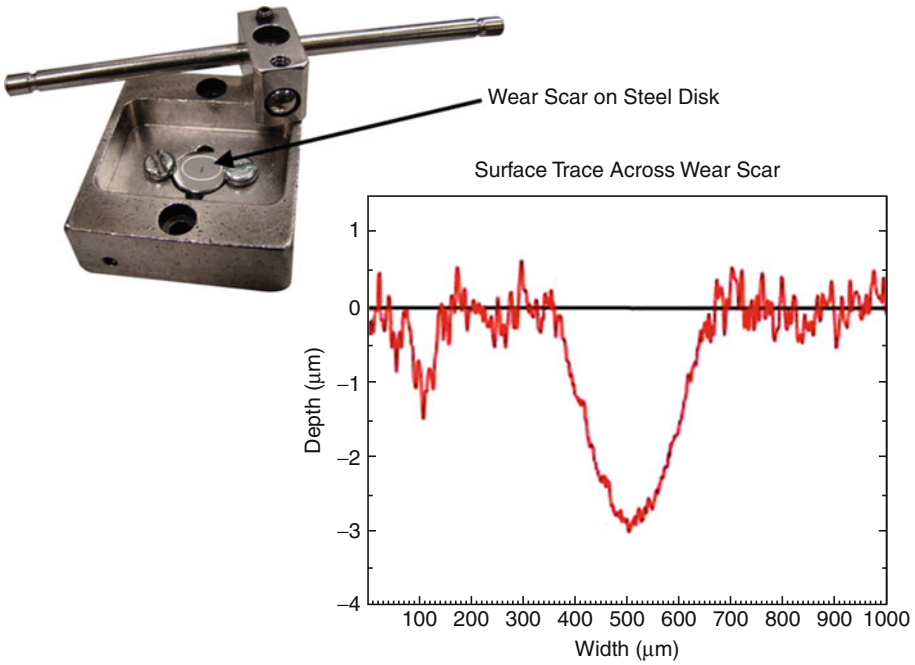
Additive Chemistry Testing Methods, Fig. 2 Scanning electron micrograph of surface fatigue damage on a spur gear (Jao et al. 2006)

a lubricant developer to more efficiently develop new lubricant additives (Rudnick 2008; Ruff 1992).

The benefits of bench tests are that they can precisely, quickly, and inexpensively measure specific aspects of a lubricant's performance. For example, Fig. 3 shows the wear scar generated on a sample specimen from a bench wear test in which a steel ball is oscillated across a steel disk. A cross-sectional trace of the wear scar is also shown in Fig. 3 and the depth of the wear scar can be determined to assess a lubricant's wear prevention properties. This bench test can measure the anti-wear performance of a lubricant in less than 1 h, at a cost of less than one thousand dollars per test. Therefore, a single lubricant can be tested multiple times to accurately determine its performance. The drawback of a bench test is that it might not adequately replicate all of the failure mechanisms that occur in a mechanical system. Therefore, more than one bench test may be needed to infer the performance of a lubricant in the mechanical system of interest.

Rig and engine tests are less precise, take longer to run, and are more expensive than bench tests. Rig and engine tests can cost between approximately ten thousand and one hundred thousand dollars and run for days if not weeks, so it may not be feasible to run multiple tests on a single lubricant. In addition, since rig and engine tests are run on full mechanical systems, there are many sources of error related to determining lubricant performance that would reduce the precision in the results from these tests. For example, lubricant temperature in an engine test is affected by the temperature of the lubricant in the sump, which can be well controlled, but also by the temperature the lubricant encounters near the combustion chamber, which is difficult if not impossible to control. The benefits of the rig and engine tests are that they use the mechanical system of interest, and some rig and engine tests are the tests required by equipment manufacturers to ensure that lubricants perform adequately. Therefore, if a lubricant performs well in a rig or engine test, no further testing of the lubricant is required. The least precise, most time-consuming, and most expensive lubricant tests are field or vehicle tests. These tests can take months or years to complete and can cost from several hundred thousand dollars to more than a million dollars because multiple mechanical systems or vehicles must be tested under a wide variety of conditions in order to determine the performance of the lubricant. The balance between relevance to actual operational performance and the precision, test time, and cost of different lubricant tests means that prudent selection of bench, rig/engine, and field tests is critical to any additive and lubricant development process.

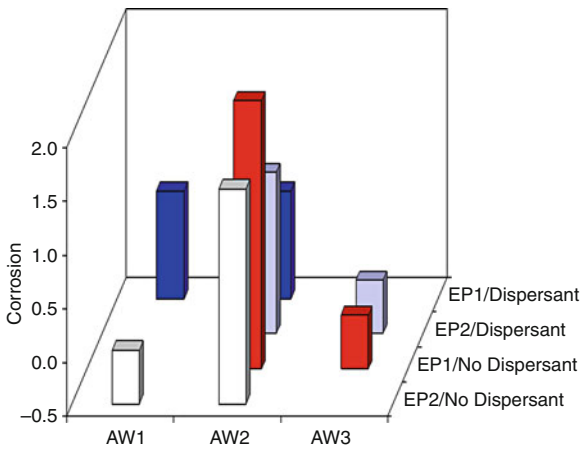
Lubricant tests must not only be designed to reflect the appropriate end-use performance, but the lubricants to be tested must be carefully selected to determine the activity of the additives present in the lubricant. As mentioned above, additive-additive and additive-base oil interactions often control overall lubricant performance. In a simple lubricant system where the properties of two additives are being evaluated in two different base oils, eight separate lubricants could be tested; each base oil by itself (two samples), each additive in each base oil (four samples), and the combination of each additive in each base oil (two samples). It is feasible to perform multiple tests on all eight lubricants in bench tests in order to gain a clear understanding of the effect of each additive and the interactions between the additives on specific lubricant properties. It is impractical because of the expense and time involved to perform multiple rig/engine or field tests on all of these same eight lubricants. In addition, since these eight lubricants may not contain the proper additives to perform all the functions of a fully formulated lubricant, the mechanical system being used in the rig/engine/field tests may not be able to operate with any of the eight lubricants. Therefore, the ability to statistically design a series of lubricants to examine additive effects and additive-additive interactions in fully formulated lubricants is critical. These lubricant sets can contain various types of additives that perform the same function (for example, different anti-wear additives) or various concentrations of the same additive. For example, Fig. 4 shows the results from DIN 51802 bearing corrosion tests performed on combinations of anti-wear additives (AW), extreme



Additive Chemistry Testing Methods, Fig. 3 Surface trace of sample specimen from a bench wear test

pressure additives (EP), and a dispersant. In the DIN 51802 test, a lubricant is mixed with salt water and the mixture is used to lubricate a roller bearing for 164 h at a speed of 80 RPM. After the test the corrosion on the bearing is rated with 0 indicating no corrosion and 5 indicating the entire surface of the bearing has been corroded. Figure 4 shows several combinations of AW, EP, and dispersant that prevent corrosion. Testing each additive separately would not have revealed the beneficial additive combinations. The statistically designed lubricant set is developed in order to minimize the number of tests required to optimize lubricant performance, which reduces testing time and cost (Myers and Montgomery 1995).

In order to develop lubricants, additive testing methods exist to measure all of the functions that a fully formulated lubricant must possess. These tests assess a lubricant's ability to (1) prevent wear and surface damage, (2) control deposits, (3) control sludge, (4) control friction, (5) resist degradation, (6) control the effects of contaminants, and (7) be transported through a mechanical system (lubricant rheology). Very precise and inexpensive tests that can be run in a chemical laboratory exist as well as more expensive and time-consuming tests that use full mechanical systems operated under controlled or actual operating conditions. The cost, time to run, and relevance to actual operational performance must be considered when selecting the tests to use to develop additives and lubricants. In addition, the sets of lubricants to be tested must be carefully designed so that additive-additive interactions are assessed and lubricant development costs and timing are reduced.



Additive Chemistry Testing Methods, Fig. 4 Corrosion rating for combinations of anti-wear additives (AW), extreme pressure additives (EP), and dispersants as measured in DIN 51802 (Devlin et al. 2003)

Key Applications

The general rationale for lubricant and additive testing is described above. This section gives examples of some bench and rig tests used to determine the performance of lubricants. The tests described in each performance area are representative of tests used to develop a wide variety of lubricants. Precise details of each test method mentioned below can be found in the American Society of Testing and Materials (ASTM) lubricant testing handbook along with many other examples of lubricant bench, rig, and engine tests. Other international organizations such as the Coordinating European Council (CEC), Deutsches Institut fur Normung (DIN), and Japan Petroleum Institute (JPI) have recommended bench, rig, and engine test methods to be used in the development of lubricants. Examples of fleet and vehicle tests are not presented because the methods are often specifically defined by equipment manufacturers and in some cases are proprietary (Rudnick 2008).

Examples of Bench Tests

1. Wear and surface damage prevention:
   ASTM D4172: Standard Test Method for Wear Preventive Characteristics of Lubricating Fluid (Four-Ball Method)

   ASTM D4172 measures the wear preventive properties of lubricating oils in a sliding contact by means of rotating a steel ball on top of three stationary balls. The load between the balls and the temperature of the lubricant can be controlled. The wear performance of the lubricant is assessed by determining the size of the wear tracks on the four balls.
2. Deposit control:
   ASTM D7097: Standard Test Method for Determination of Moderately High Temperature Piston Deposits by Thermo-Oxidation Engine Oil Simulation Test—TEOST MHT

   ASTM D7097 measures the deposit-forming tendencies of engine lubricants. The amount of deposit formed on a test rod exposed to repetitive passage of engine oil over the rod is determined. The oil and rod are exposed to high temperatures, causing accelerated oxidation of the oil and deposit formation on the rod.
3. Sludge control:
   ASTM D5763: Standard Test Method for Oxidation and Thermal Stability Characteristics of Gear Oils Using Universal Glassware

   ASTM D5763 measures the ability of gear oils to resist thermal and oxidative degradation that can result in the formation of sludge. The gear oil is exposed to an oxidative environment at high temperature in glassware. At the end of the test the amount of sludge, viscosity change, and oil loss are determined.
4. Friction control:
   ASTM D6425: Standard Test Method for Measuring Friction and Wear Properties of Extreme Pressure (EP) Lubricating Oils Using SRV Test Machine

   ASTM D6425 measures the coefficient of friction of lubricating oils under high sliding conditions or start-stop motion. The load between the surfaces in contact in this method and the temperature of the lubricant can be controlled.
5. Lubricant degradation:
   ASTM D6186: Standard Test Method for Oxidation Induction Time of Lubricating Oils by Pressure Differential Scanning Calorimetry (PDSC)

   ASTM D6186 measures the oxidative stability of lubricants by exposing the lubricant to high temperatures and high oxygen pressure. Once the lubricant is exposed to the oxidizing conditions, the time until an exothermic oxidation reaction is measured in order to determine the oxidative stability of the lubricant.
6. Contaminant effects on lubricant performance:
   ASTM D1401: Standard Test Method for Water Separability of Petroleum Oils and Synthetic Fluids

   ASTM D1401 measures the water separation characteristics of lubricants. A precise amount of water is blended into a lubricant and the time until the formation of separate water and lubricant phases is measured. If complete separation does not occur, the relative amounts of water, lubricant, and emulsion phase are determined. This method can be performed at various temperatures.
7. Lubricant rheology
   ASTM D2983: Standard Test Method for Low-Temperature Viscosity of Lubricants Measured by Brookfield Viscometer

   ASTM D2983 measures the low-temperature, low-shear-rate viscosity of gear oils, automatic transmission fluids, tractor fluids, and hydraulic fluids. A lubricant sample is cooled to a test temperature between 5°C and −40°C. Then a motor rotates a spindle within the sample and the torque required to rotate the spindle is recorded and used to calculate the viscosity of the lubricant.

Examples of Rig/Engine Tests

1. Wear and surface damage prevention:
   ASTM D6121: Standard Test Method for Evaluation of Load-Carrying Capacity of Lubricants Under Conditions of Low Speed and High Torque Used for Final Hypoid Drive Axles

ASTM D6121 measures a gear lubricant's ability to prevent wear and surface fatigue in hypoid gears operating under low-speed, high-torque conditions.

2. Deposit control:
   ASTM D7320: Standard Test Method for Evaluation of Automotive Engine Oils in the Sequence IIIG, Spark-Ignition Engine
   ASTM D7320 measures an engine lubricant's ability to control varnish deposition, oil thickening, oil consumption, and wear in an engine operating at high speeds and high temperatures.
3. Sludge control:
   ASTM D6593: Standard Test Method for Evaluation of Automotive Engine Oils for Inhibition of Deposit Formation in a Spark-Ignition Internal Combustion Engine Fueled with Gasoline and Operated Under Low-Temperature, Light-Duty Conditions
   ASTM D6593 measures an engine lubricant's ability to control sludge formation and piston deposits in an engine that is cycled from high speeds and high temperatures to low speeds and low temperatures. These operating conditions are selected to accelerate the formation of sludge and deposits.
4. Friction control:
   ASTM D6837: Standard Test Method for Measurement of Effects of Automotive Engine Oils on Fuel Economy of Passenger Cars and Light-Duty Trucks in Sequence VIB Spark Ignition Engine
   ASTM D6837 measures an engine lubricant's ability to improve vehicle fuel economy. Fuel consumed while operating an engine with a selected lubricant is measured under test conditions similar to those encountered in vehicles during normal city and highway driving.
5. Lubricant degradation:
   ASTM D5704: Standard Test Method for Evaluation of the Thermal and Oxidative Stability of Lubricating Oils Used for Manual Transmissions and Final Drive Axles
   ASTM D5704 measures the tendency of automotive manual transmission and final drive lubricants to deteriorate under high-temperature conditions. The ability of the lubricants to prevent oil-thickening, insolubles-formation, and deposit-formation are determined under high-temperature oxidizing conditions.
6. Contaminant effects on lubricant performance:
   ASTM D7156: Standard Test Method for Evaluation of Diesel Engine Oils in the T-11 Exhaust Gas Recirculation Diesel Engine
   ASTM D7156 measures the ability of engine lubricants to control soot-induced viscosity increase in diesel engines equipped with exhaust gas recirculation devices.
7. Lubricant rheology:
   The bench test methods that are used to measure the rheological properties of lubricants are often used to measure the rheological properties of lubricants that have been aged in rig/engine and fleet tests.

Cross-References

- ▶ Corrosive Wear
- ▶ Engine Lubricants
- ▶ Engine Oil Test Equipment
- ▶ Fatigue
- ▶ Friction Measurement
- ▶ Gear Lubricants
- ▶ Lubricant Formulation
- ▶ Lubricant Viscosity
- ▶ Rheology – Viscosity Index
- ▶ Sliding Wear
- ▶ Transmission Lubricants
- ▶ Tribochemistry of Antiwear (AW) Additives
- ▶ Wear in Gears
- ▶ Wear of Bearings

References

M.T. Devlin, H. Ryan, V. Tsang, P. Corbett, L. Strand, T.L. Turner, C. Wallo, T.-C. Jao, The effect of water contamination and oxidation on the fatigue life performance of wind turbine lubricants. Presented at the NLGI 70th Annual Meeting. Hilton Head, SC, 2003

R.F. Haycock, J.E. Hillier, A.J. Caines, *Automotive Lubricants Reference Book*, 2nd edn. (SAE International, Warrendale, 2004)

T-C. Jao, M.T. Devlin, J. Milner, R. Iyer, M.R. Hoeprich, Influence of surface roughness on gear pitting behavior. Gear Technol. (May/June 2006), pp. 30–38

R.H. Myers, D.C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (Wiley, New York, 1995)

L.R. Rudnick, Testing methods for additive/lubricant performance, in *Lubricant Additives: Chemistry and Applications*, ed. by L.R. Rudnick, 2nd edn. (CRC Press, Boca Raton, 2008), p. 669

A.W. Ruff, in *Laboratory Characterization Techniques, Friction, Lubrication, and Wear Technology, ASM Handbook*, ed. by P.J. Blau, S.D. Henry, vol. 18 (ASM International, Materials Park, Ohio, 1992), p. 333

Adhesion and Electrostatic Field

▶ Electrostatic Field Effects on Adhesion

Adhesion and Lubricious Coatings

▶ Solid–Liquid Bi-phase Lubricating Coatings

A

Adhesion Hysteresis

Linmao Qian, Bingjun Yu
Tribology Research Institute, National Traction Power Laboratory, Southwest Jiaotong University, Chengdu, Sichuan Province, People's Republic of China

Definition

Adhesion hysteresis describes the phenomenon where taking apart two contact surfaces dissipates more energy than bringing the surfaces together.

Scientific Fundamentals

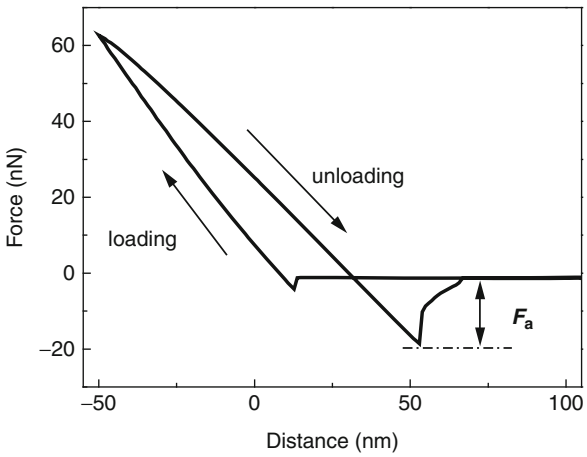
The adhesion phenomenon exists widely in atomic force microscope (AFM)-based nanotribology studies (Qian and Xiao 2000). As shown in Fig. 1, when the AFM tip contacts to the silicon surface (loading) and then pulls off (unloading) from the sample, the adhesive force F_a can be obtained from the force-distance curve. The surface energy has a positive effect on F_a.

Theoretically, the work of adhesion W or surface tension (surface energy) γ, where $W = 2\gamma$, is normally viewed as the reversible work done on bringing two surfaces together or the work needed to separate two surfaces from contact. But under most realistic conditions, it will dissipate more energy to take apart two contact surfaces than bring two surfaces together, which can be defined as *adhesion hysteresis* (Bhushan 2001). Further understanding of the molecular mechanisms underlying this phenomenon is essential for explaining many adhesion phenomena, energy dissipation during loading-unloading cycles, contact angle hysteresis, and the molecular mechanisms associated with many frictional processes.

Adhesion hysteresis may be thought of as being due to mechanical or chemical effects, namely the physical and chemical changes process is irreversible during the contact-separating process. In general, if the energy change, or work done, on separating two surfaces from adhesive contact is not fully recoverable on bringing the two surfaces back into contact again, the adhesion hysteresis may be expressed as (Chen et al. 1991)

$$\begin{array}{ccc} W_R & > & W_A \\ \text{receding} & & \text{advancing} \\ \text{(separating)} & & \text{(approaching)} \end{array} \quad \text{or} \quad \Delta W = W_R - W_A > 0 \tag{1}$$

where W_R and W_A are the adhesion or surface energies for receding (separating) and advancing (approaching) solid surfaces, respectively.



Adhesion Hysteresis, Fig. 1 Force-distance curve of approaching and separating process

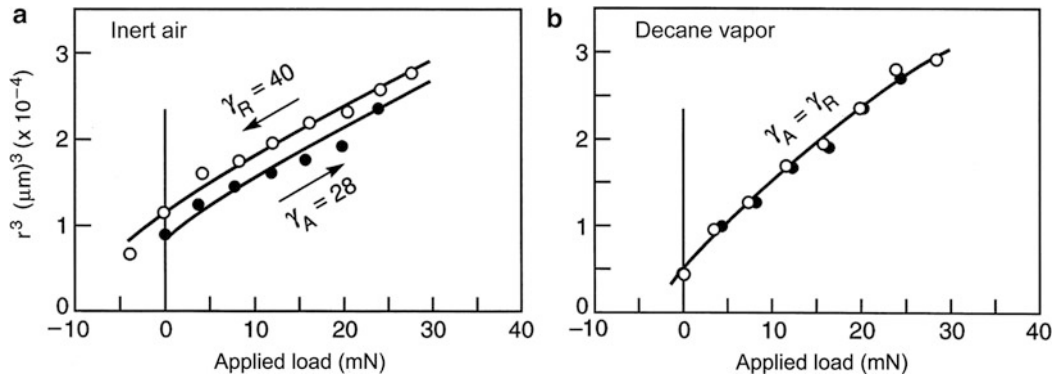
Adhesive hysteresis can also be found in measurement of contact angles. For example, when the liquid spreads and then retracts from a surface, the advancing contact angle θ_A is generally larger than the receding angle θ_R (Bistac et al. 1998). The contact angle θ is related to the liquid-vapor surface tension γ_L and the solid-liquid adhesion energy W by the following equation:

$$(1 + \cos\theta)\gamma_L = W \tag{2}$$

It can be concluded that wetting hysteresis or contact angle hysteresis ($\theta_A > \theta_R$) implies that either $\gamma_{L,A} > \gamma_{L,R}$ or $W_R > W_A$. The wetting hysteresis or contact angle hysteresis actually implies the adhesion hysteresis as (1) presents.

Molecular dynamics (MD) simulations (Landman et al. 1990) have predicted hysteresis in the force versus tip-sample distance related to intrinsic mechanical instabilities at the atomic scale. As the tip approaches the sample, the interaction is essentially given by the attractive conservative forces (i.e., the hysteretic force of adhesion is equal to zero). Just before contact, there is a sudden jump of the interaction force due to the formation of an atomic scale connective neck and, as the tip retracts, there is an additional adhesive force that drops approximately linearly in a few interatomic distances (D_0). This behavior associated with the formation and rupture of a solid neck (Landman et al. 1990) is similar (except for some oscillations due to atomic rearrangements) to the capillary-induced liquid bridges (Sahagún et al. 2007). So a linear adhesive force (when the tip retracts) can be simply estimated (Köber et al. 2008) for $D_{ts} < D_0$ by

$$F_{hys} \approx \frac{2\Delta E}{D_0^2}(D_{ts} - D_0) \tag{3}$$



Adhesion Hysteresis, Fig. 2 Contact radius vs. applied load (r^3–L) curves measured from two fluid-like monolayer-coated surfaces in inert air (**a**) and decane vapor (**b**). The *solid lines* are obtained from the analysis of these data using the JKR equation

$$r^3 = \frac{R}{K}\left[F + 6\pi R\gamma + \sqrt{12\pi R\gamma F + (6\pi R\gamma)^2}\right] \qquad (4)$$

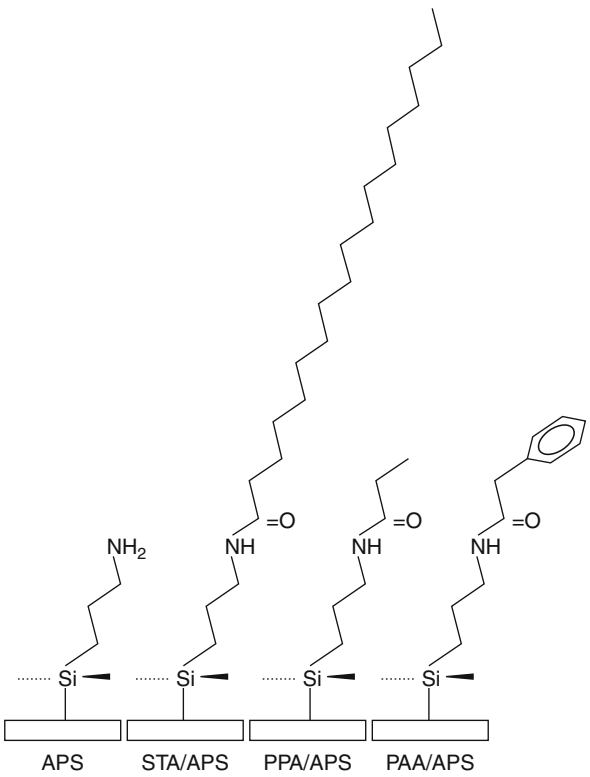
where $R = R_1R_2/(R_1 + R_2)$ (R_1 and R_2 stand for the radii of two spheres), K is bulk elastic moduli, and γ is surface energy. For dry monolayers (**a**) the adhesion energy on unloading (γ_R = 40 mJ/m^2) is greater than that on loading (γ_A = 28 mJ/m^2), and the adhesion energy hysteresis can be calculated as $\Delta\gamma = \gamma_R - \gamma_A$ = 12 mJ/m^2. For monolayers exposed to saturated decane vapor (**b**), the adhesion hysteresis disappeared (From (Chen et al. 1991). With permission)

where F_{hys} is the hysteretic force of adhesion and ΔE is the energy dissipated in the contact process.

Energy dissipating processes originate from practical constraints of the finite time of measurements and the finite elasticity of materials, which prevent many loading-unloading or approach-separation cycles from being thermodynamically reversible (Bhushan 2001).

Key Applications

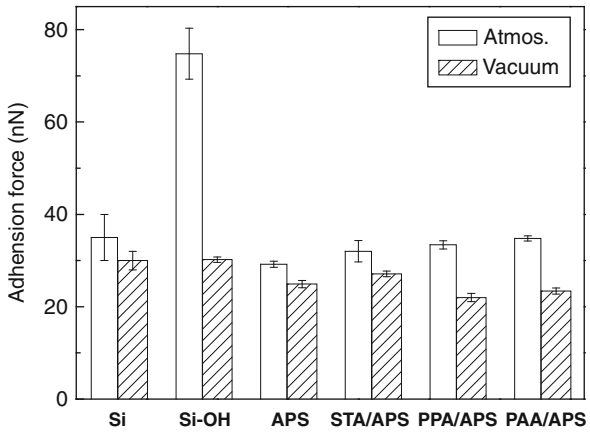
Silicon-based micro/nanoelectromechanical systems (MEMS/NEMS) experience oxidation during micromachining and subsequent exposure to air, forming a surface of increased stiction and friction (Chandross et al. 2004). In these applications, the surface forces, such as friction and adhesion, play a much more important role than the bulk forces because of the surface and size effects in nanoscale. For instance, the adhesion has induced the negative influence on the efficiency and reliability of the digital micro-mirror devices (Bhushan and Liu 2004). Therefore, with the development in MEMS/NEMS, the understanding and control of the adhesion and friction performance have become an important issue of concern. Further study on adhesion hysteresis can help in understanding the correlation between adhesion and friction performance in real application of MEMS/NEMS, which can shed new light on the design of demanded operating surface of devices.



Adhesion Hysteresis, Fig. 3 Schematic structures of self-assembled films on Si(111) substrate (Figs. 3–5: From Yu et al. 2009. With permission)

Figure 2 shows that the adhesion energy hysteresis ($\Delta\gamma$) has a direct impact on the friction between two interactional surfaces in movement, namely two surfaces with high adhesion hysteresis will lead to a high friction during movement. On the other hand, surface with low surface energy or adhesion will produce less adhesion hysteresis and thus low friction will be obtained in a nanoscale test. This is of significance for contact surfaces of MENS/NEMS, which can easily suffer from friction and adhesion in operations.

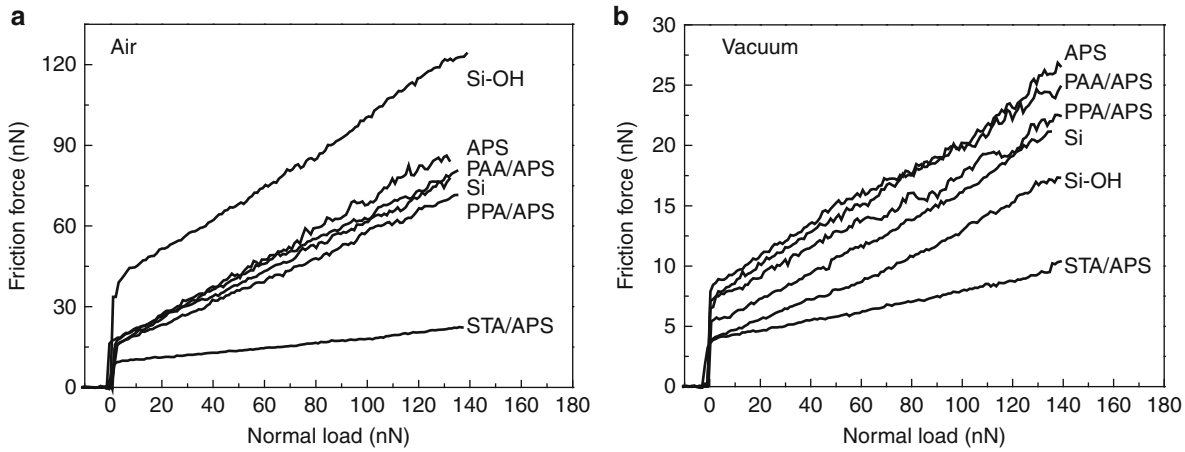
In practice, a modified surface with low adhesion usually exhibits low friction and good antiwear performance. As shown in Fig. 3 (Yu et al. 2009), dual-layer films of STA/APS, PPA/APS, and PAA/APS were grafted to the Si(111) surface. The adhesion force between the AFM tip and sample surface was measured and recorded in Fig. 4. It was found that for all samples, the adhesion force tested in air was higher than that in vacuum, especially for the hydrophilic Si–OH surface. The hydrophilic Si–OH surface showed the largest variation in its adhesion force, which decreased from 74.8 nN in air to 30.2 nN in vacuum. This is because, in air, water meniscus may form on samples depending on their hydrophilicity and humidity, which may further induce an additional capillary adhesion between AFM tip and sample (Qian et al. 2003).



Adhesion Hysteresis, Fig. 4 Adhesion forces of Si with native surface, Si–OH surface, and self-assembled films measured in air and in vacuum (Yu et al. 2009)

Considering that the adhesion of samples was smaller when measured in vacuum than in air, it can be inferred that these films will exhibit accordingly lower friction in a vacuum test based on preceding discussions of adhesion hysteresis. Figure 5 illustrates the friction force versus normal force curves of Si, Si–OH, and self-assembled films measured in air and in vacuum. The dual-layer film STA/APS constructed by densely packed long chains revealed much lower friction force than PPA/APS film by poor-packed short chains in both air and vacuum (Fig. 5a, b) (Xiao et al. 1996). Compared with the PPA/APS film with the tail methyl groups, the PAA/APS film presented a relatively high friction force resulting from the distortion and rotation of phenyl groups.

In summary, a comprehensive understanding the adhesion hysteresis can help us predict the magnitude of friction force. Minimizing adhesion hysteresis can effectively improve the micro- and nanotribological performance of contact surfaces during operation, which is practical in designing new contact surfaces, such as self-assembled film-coated surfaces, of MENS/NEMS devices.



Adhesion Hysteresis, Fig. 5 Friction force versus the applied load curves of Si, Si–OH surface, and self-assembled films measured in (**a**) air and (**b**) vacuum (Yu et al. 2009)

Cross-References

- ► Basic Concepts in Adhesion Science
- ► Interfacial Energy
- ► Liquid Contact Angle Measurement
- ► Self-Assembled Monolayers
- ► Surface Forces, Surface Tension, and Adhesion
- ► Surface Free Energy

References

B. Bhushan, *Modern Tribology Handbook*, vol. 2 (CRC Press LLC, Boca Raton, 2001)

B. Bhushan, H. Liu, Micro/nanoscale tribological and mechanical characterization for MEMS/NEMS. Proc. SPIE **5392**, 1–13 (2004)

S. Bistac, P. Kunemann, J. Schultz, Tentative correlation between contact angle hysteresis and adhesive performance. J. Colloid Interface Sci. **201**, 247–249 (1998)

M. Chandross, E.B. Webb III, M.J. Stevens, G.S. Grest, Systematic study of the effect of disorder on nanotribology of self-assembled monolayers. Phys. Rev. Lett. 93(16) (2004), Art. No. 166103

Y.L. Chen, C.A. Helm, J.N. Israelachvili, Molecular mechanisms associated with adhesion and contact angle hysteresis of monolayer surfaces. J. Phys. Chem. **95**, 10736–10747 (1991)

M. Köber, E. Sahagún, M. Fuss, F. Briones, M. Luna, J.J. Sáenz, Adhesion hysteresis in dynamic atomic force microscopy. Phys. Stat. Sol. (RRL) **2**(3), 138–140 (2008)

U. Landman, W.D. Luedtke, N.A. Burnham, R.J. Colton, Atomistic mechanisms and dynamics of adhesion, nanoindentation, and fracture. Science **248**, 454–461 (1990)

L.M. Qian, X.D. Xiao, Tip in situ chemical modification and its effects on tribological measurements. Langmuir **16**, 662–670 (2000)

L.M. Qian, F. Tian, X.D. Xiao, Tribological properties of self-assembled monolayers and their substrates under various humid environments. Tribol. Lett. **15**(3), 169–176 (2003)

E. Sahagún, P. García-Mochales, G. M. Sacha, J.J. Sáenz, Energy dissipation due to capillary interactions: hydrophobicity maps in force microscopy. Phys. Rev. Lett. 98 (2007), Art. No. 176106

X.D. Xiao, J. Hu, D.H. Charych, M. Salmeron, Chain length dependence of the frictional properties of alkylsilane molecules self-assembled on mica studied by atomic force microscopy. Langmuir **12**, 235–237 (1996)

B.J. Yu, L.M. Qian, J.X. Yu, Z.R. Zhou, Effects of tail group and chain length on the tribological behaviors of self-assembled dual-layer films in atmosphere and in vacuum. Tribol. Lett. **34**, 1–10 (2009)

Adhesion in the Animal World

Zhendong Dai
Institute of Bio-inspired Structure and Surface Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, People's Republic of China

Synonyms

Attractive interaction force; Capillary force; Interfacial energy in biosystems; Surface forces in biosystems; Surface tension in biosystems; van der waal forces in biosystems

Definition

Adhesion is defined as the physical attraction or joining of two substances, especially the macroscopically observable attraction of dissimilar substances, or as the force that holds together the molecules of unlike substances whose surfaces are in contact.

Scientific Fundamentals

Adhesive Device in Animals

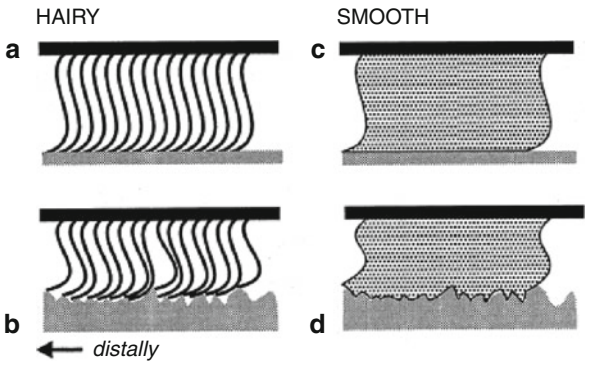
To escape from predators, to find food, and to search for suitable mates, animals have developed abilities to overcome the forces of gravity and inertia on their bodies and of unexpected exterior forces acting on their bodies (Dickinson et al. 2000). Reliable attachment between foot and substrate is most important for animals moving on the land, especially when moving on substrates with large slope angles (such as $\geq 90^{\circ}$) to the horizontal surface. In such cases, animals must find a way to generate an attraction force between their feet and the substrate in order to overcome the gravity acting on the animal's body, rather than the repulsive force between contacted surfaces. This attraction force is called adhesion.

Over thousands of years of evolution, animals have optimized claws, smooth pads, and hairy pads to hold them more stable. The interaction of claw with substrate is determined by roughness of substrate, the friction coefficient between claw and substrate, and the relative dimension of claw and the substrate, which forms frictional self-locking (Dai et al. 2002). Both smooth pads and hairy pads may generate adhesion, but the mechanism is thought of as capillary force and van der Waals force, respectively, and the adhesion is generally considered as wet adhesion and dry adhesion. The structures of both pads aim to maximize the possible real contact area with the substrate, regardless of its micro-sculpture (Fig. 1) (Beutel and Gorb 2001).

The adhesive devices in various animals are located differently and many animals have more than one adhesive device. The attachment devices in most hexapods is located on different parts, such as claws, derivatives of the pretarsus, tarsal apex, tarsomeres, or tibia of insects (Fig. 2) (Beutel and Gorb 2001). The adhesive pads are adapted for holding onto smooth plant substrates where claws fail to get a grip. The adhesive devices of insects show the diversity on the morphology, location, and the geometrics scale, but they are basically smooth pads or hairy pads.

Geckos possess an excellent ability to move on various substrates. The structure of the gecko foot, the lamellae or seta array and the micro-structure of setae, were

extensively studied (Autumn et al. 2000). The hairy adhesion strongly depends on the distance between terminal part of hair and the substrate. The density of the seta and the diameter of the seta are also a very important parameter (Fig. 3), with increases of body weight in animals, the diameter of the seta decreases to generate enough adhesion to balance the gravity of the animals, that is, with increase of body weight of the animals, the scale of their adhesive seta decreases (Arzt et al. 2003).

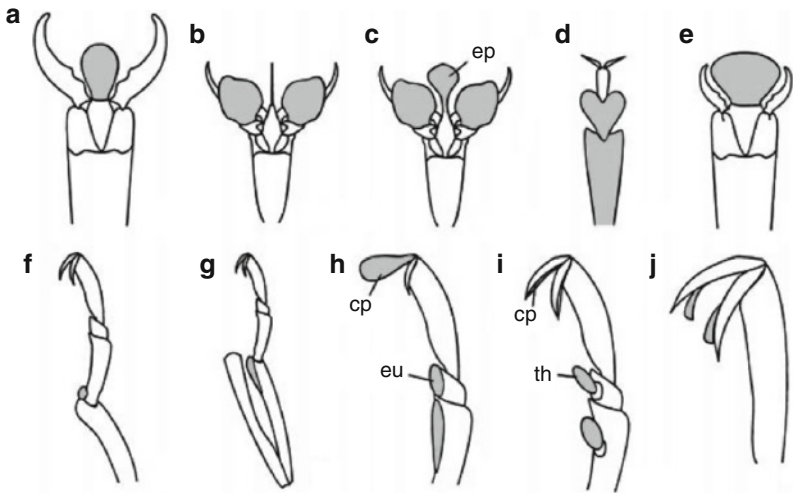


Adhesion in the Animal World, Fig. 1 Illustration on the adhesion of hairy (**a**, **b**) and smooth (**c**, **d**) pads on smooth (**a**, **c**) and rough (**b**, **d**) substrate. Both structures are able to adapt the substrate profile to maximize the real contact area

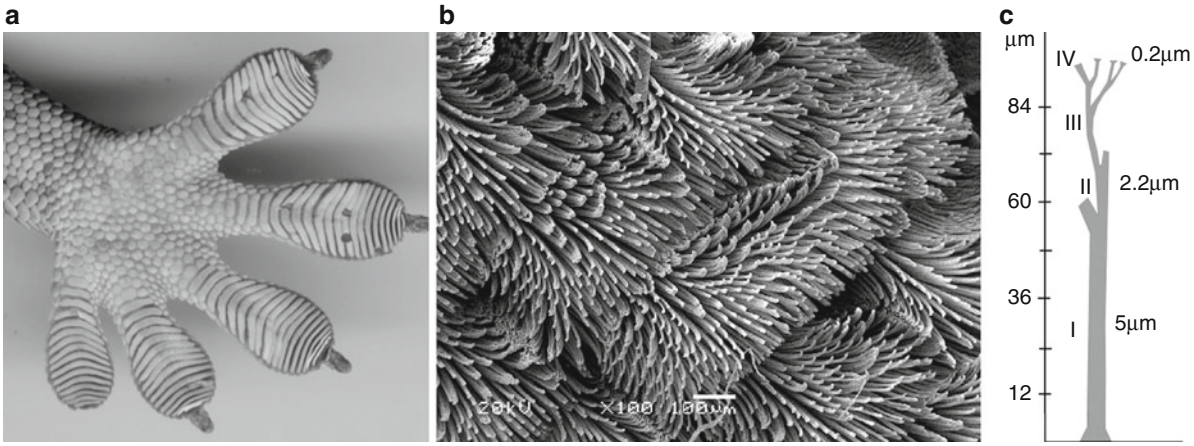
Smooth and deformable cuticle pads were found in many insects. During attachment, the contacted surface is mediated by adhesive liquid (Jiao et al. 2000), the secretion appears to be released onto the surface of the pads through pore canals or channels, but its exact pathway still need to be explored.

Tree frogs are remarkable for their capacity to cling to smooth surfaces using large toe pads. The adhesive skin of toe pads is characterized by hexagonal cells separated by deep channels into which mucus glands open. The pads are completely wetted with watery mucus, which show that attachment is solely due to capillary and viscous forces (Fig. 4). Tree frogs have less compliant pad surface layers, and in these cases adhesion to rough surfaces is only possible because the animals inject a wetting liquid into the pad–substrate contact area, which generates a relative long-range attractive interaction due to the formation of capillary bridges (Persson 2007).

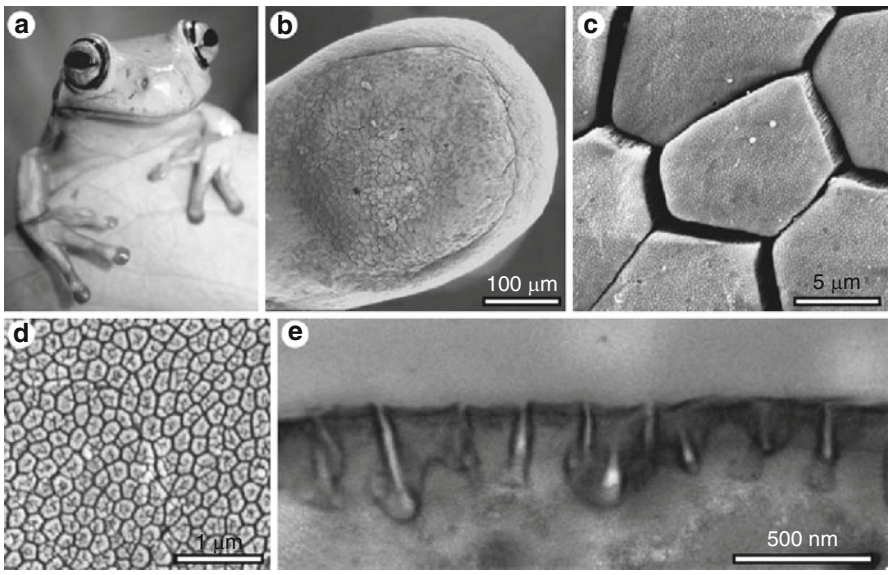
Attachment on a smooth surface for 11 ant species with different wax-running capacities was compared (Federle et al. 2000). The results provide evidence that superior "wax-runners" did not cling better to smooth Perspex and performed significantly worse than closely related congeners that are unable to climb up waxy stems. This suggests an inverse relationship between adaptations to run on wax and to attach to a smooth surface.



Adhesion in the Animal World, Fig. 2 Diversity of the leg attachment devices (*gray* areas) in insects. (**a**) Smooth arolium; (**b**) Smooth or hairy pulvilli; (**c**) Hairy empodial pulvilli (*ep*); (**d**) Hairy adhesive soles of tarsomeres; (**e**) Smooth eversible pretarsal bladder; (**f**) Smooth eversible structure between tibia and tarsus; (**g**) Hairy fossula spongiosa; (**h**) Smooth euplantulae (*eu*) and claw pad (*cp*); (**i**) Smooth tarsal thoms transformed into adhesive structures (*th*) and claw pad (*cp*); (**j**) Adhesive claw pad (After Beutel and Gorb 2001)



Adhesion in the Animal World, Fig. 3 The structure of gecko foot and the micro- structure of setae (**a**) gecko foot structure; (**b**) Lamellae on the toe; (**c**) microstructure of a single setae

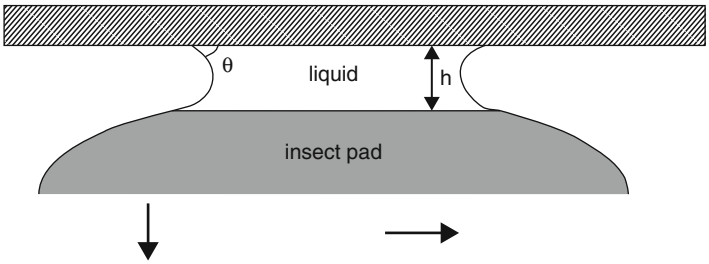


Adhesion in the Animal World, Fig. 4 Morphology of tree frog toe pads. (**a**) White's tree frog (*Litoria caerulea*). (**b**–**d**) SEMs of (**b**) toe pad, (**c**) epidermis with hexagonal epithelial cells and (**d**) high-power view of the surface of a single hexagonal cell showing peg-like projections. (**e**) TEM of cross-section through cell surface

Adhesive organs on the legs, such as in cockroaches, are strongly direction dependent, making contact only when pulled towards the body but detaching when pushed away from it. Single-leg friction force measurements showed that the arolium and euplantulae have an opposite direction dependence. Euplantulae achieved maximum friction when pushed distally, whereas arolium forces were maximal during proximal pulls. The results suggest that the euplantulae in cockroaches are not adhesive organs but "friction pads," mainly providing the necessary traction during locomotion (Clemente and Federle 2008).

Wet Adhesion Mechanism

Wet adhesive pads were found in most smooth soft pads and some hairy pads (Huber et al. 2005), where a liquid film between pads of animals and the substrate gives rise to adhesive forces due to surface tension and viscosity (Fig. 5). On the direction perpendicular to the surface,



	vertical force (adhesion)	parallel force (friction)	θ : contact angle h : liquid film height γ_L : surface tension η : viscosity v : velocity R : pad radius of curvature A : pad area
static	$F = 3\pi R \gamma_L \cos\theta$ (Fogden & White 1990)	*Lubrication: low friction expected*	
dynamic	*additional viscosity contribution*	$F = \frac{A}{h} \cdot v \cdot \eta$	

Adhesion in the Animal World, Fig. 5 Illustration of the wet adhesive contact and prediction from the wet adhesive mechanism. Static force was predicted on the hypothesis of soft sphere and the dynamic force was predicted on the parallel plate with Newtonian liquid (Federle et al. 2002)

static forces are created mainly by surface tension, but the dynamic forces are generated by the surface tension and viscosity of the liquid by the relative motion; On the other hand, forces parallel to the surface can be negligible for static attachment, but they are quite big when relative motion happens due to very small distance between the surfaces. Several predictions for insect attachment forces follow from these considerations (Federle et al. 2002).

1. Static friction should be small, but dynamic friction should be much larger. The character may explain why the frictional forces are much larger than adhesive forces.
2. Friction should depend on velocity and the viscosity of the liquid. Due to shearing of the liquid film, the friction force should be stronger at higher sliding velocities.
3. The effects of temperature on viscosity and surface tension are quite different. The dependence of sliding friction should become smaller at higher temperatures, but static forces should be almost temperature-independent.

Figure 5 suggests that contact area, distance between pad and substrate, viscosity, and surface tension of the liquid filling into the distance are the most important factors that determine the adhesion and friction of soft smooth pad contact. The contact area depends on the force acting on pad and the stiffness of the pad because the substrate, in most cases, is very hard compared with the smooth soft pads. During the attaching procedure, the force that acts on the pad at the beginning has to be balanced by the adhesion acting on other pads, so animals tend to minimize the force. On the other hand, the lower stiffness of the pad will dramatically increase the contact area. On a rough surface, animals have developed a technique to separate the surface of soft pad into many smaller areas to adapt the substrate profile and decrease the distance between two surfaces. The third factor is the performance of liquid secreted by animals. Viscosity of the liquid, wettability of the liquid to the substrate, and the surface tension will determine the adhesion in the normal direction and the friction in parallel to the contact surface.

A liquid film between two objects gives rise to adhesive forces due to surface tension and viscosity. Surface tension mainly creates static forces perpendicular to the surface. Forces due to viscosity can act both in the normal and in the parallel direction, but they are zero in the static case.

To understand the mechanism of wet adhesion, a model was proposed to explain the force between pad and surface (Fig. 5), the definite liquid would be contracted or expanded when insect pad was pulled off or pushed. It is supposed that γ_L is the surface tension, h is the height of liquid, the radius of contact area between liquid and pad R, the contact angles are θ, so the adhesion force is presented as:

$$F = 3\pi R \gamma_L \cos\theta \qquad (1)$$

The adhesive properties are significantly affected by density of the fibers, thickness of the superficial layer,

and compliance of the pad. Grasshoppers and locusts have a similar structure of their attachment pads. The morphology, ultra-structure, effective elastic modulus, and adhesive properties of these two different smooth-type attachment pads are somewhat different. Locusts' pads bear a thicker sub-superficial layer and a higher density of rods. Indentation experiments showed a higher effective elastic modulus and a lower work of adhesion for locust pads.

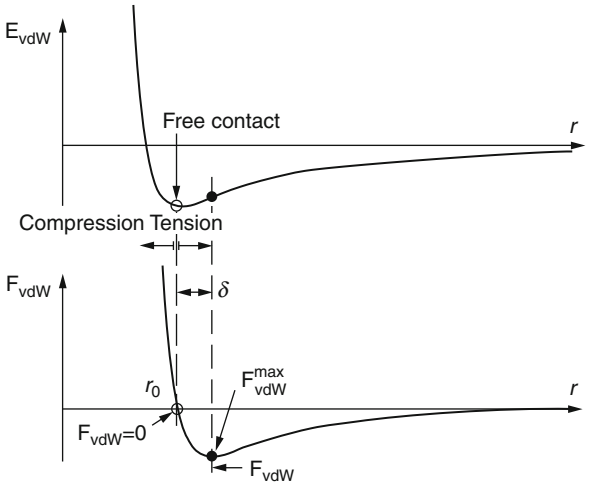
Dry Adhesion Mechanism

Dry adhesion is geometrically related with hairy pads, which occurs in many insects (such as beetles), reptiles (such as geckos), and spiders. It was demonstrated that van der Waals interaction between molecular plays a key role in the adhesion. Research shows the distance between the terminal of the hair on the pads and the substrate is much smaller than that in soft smooth pads. Among them, geckos' extraordinary climbing ability has attracted much attention and study over the past years. The gecko is considered a remarkable design of nature that is attributed to the fine structure of its toes, which contain setae arrays consisting of hundreds of spatulae on each seta. Although micrometer dimensions of the terminal elements of the setae are sufficient for flies and beetles, geckos must resort to sub- micrometer devices to ensure adhesion (Arzt et al. 2003).

The adhesive force of a single seta was measured (Autumn et al. 2000) and the adhesion mechanism was believed as van der Waals only for the hairy system (Autumn et al. 2002). This attractive force exists in any non-polar object. It is caused by fluctuations in the instantaneous dipole moments of these two objects due to the uneven distribution of electrons in their electron clouds. Two adjacent particles tend to synchronize their dipole moment fluctuations to minimize the total potential energy; therefore van der Waals forces are usually attractive and are the smallest among all intermolecular forces, but they become significant when a large number of particles are involved in suitable distance (very small). The attraction between two solids can be calculated by integrating the London dispersion energy over all particles in both volumes, and differentiating with respect to the separation distance between them.

From microscope point view, the potential E_{vdW} of an intermolecular pair is obtained by summing the attractive potential E_A and repulsive potential E_R, and the best known form is the Lennard-Jones equation, where m and n are 12 and 6, respectively (Israelachvili 1985)

$$E_{vdW} = -E_A(r/r_0)^{-n} + E_R(r/r_0)^{-m} = 4\varepsilon[(r_0/r)^{12} - (r_0/r)^6] \quad (2)$$



Adhesion in the Animal World, Fig. 6 The Lennard-Jones potential and force between two particles

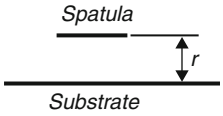
Where ε is a constant of interaction potential. The intermolecular potential and force are illustrated in Fig. 6.

The illustration shows the relation of potential and force to the distance between particles. When the distance between two particles $r = r_0$, the repulsive force equals the attractive force, the potential of the system becomes minimum. When two particles are pressed to $r < r_0$, the molecular interaction F_{vdW} is repulsive, which will dramatically increase with decrease of distance *r*. When the distance between two particles falls into $r > r_0$, the interaction becomes attractive and increases with increase of distance until $r_0 < r < r_0 + \delta$, then, the attractive force decreases with increase of distance, so the force reaches its maximum F_{vdW}^{max} (the "adhesion" or "pull-off" force) at distance $r = r_0 + \delta$. When $r > r_0 + \delta$, the interaction is still attractive, but it decreases with increase of distance, so the attractive link will be soon broken. The summed force of setae array is an integration of repulsive forces and attractive forces of each hair on the array. To obtain great adhesive force, animals have evolved techniques to reduce the repulsive force and increase the attractive force by inclining the seta to decrease the contact stiffness of seta to the substrate.

Supposing the terminal part of seta and the substrate are two parallel surfaces (Fig. 7), the adhesive force to a surface can be roughly modeled for the configuration per unit contact area

$$F_{vdw} = A/(6\pi r_0^3) \quad (3)$$

where A is the Hamaker constant that depends on the materials of the two surfaces, the typical value is on the order of 10^{-19}J between two solids, and does not vary



Adhesion in the Animal World, Fig. 7 Spatula and surface approximated as two parallel surfaces in contact

significantly for different materials. Note that there is a significant difference between real and apparent contact areas. Solid surfaces are rarely ideally planar. Therefore, the real contact area is merely the total of the areas between the few opposing asperities actually in position to touch each other.

Key Applications

Locust's Smooth Pads Inspired the Foot of Legged Robot

The legged robot and locust would face very similar problems during motion, reducing the impact force when foot transfer from swing phase to stance phase, making a reliable attachment when foot under stance phase. Studying the macro-structure and the material topology of wet adhesive pads and revealing mechanical roles by using finite element method may inspire engineers in developing new feet for legged robot (Dai and Gorb 2009). They studied the contact mechanics, stiffness, friction force generated on the contacted areas, and the restrained forces on the pad. The microstructure of the pads is composed of two layers: a superficial density soft layer EXO (6–10 μm) and the grove-like structures under the EXO. The rods are 2–3 μm in diameter and 40–50 μm in length, which are chitin filaments, perpendicular to the primary surface and rarely link each other (Fig. 8), so only EXO is included in FEM model (Fig. 9).

A simplified two-dimensional model is shown in Fig. 9b. Where airbag (AS) and tendon (TD) were excluded, the geometric model was exactly fit from the SEM image of the cross-section of the third tarsomere. The FEM model consists of three parts – the soft layer exocuticle (EXO), the hard cuticle (HK), and the liquid-filled container, and it is called the "fluid-contained" (FC) model. For comparison, another model consisting of only the EXO and HK was set up and named as the solid soft-material (SS) model. The two models are shown in Fig. 9b and a, respectively.

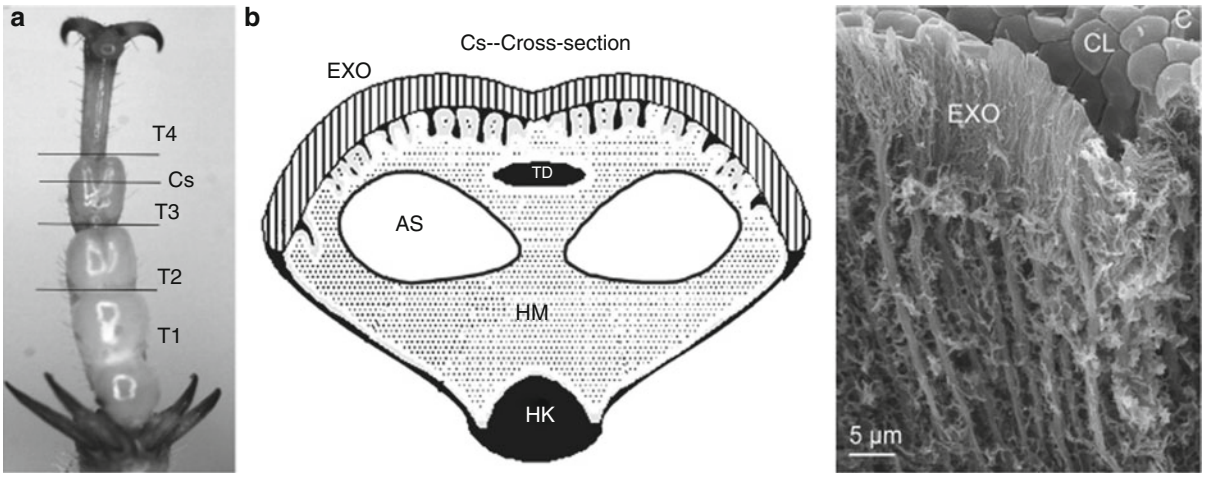
The elastic modules of the biomaterials EXO, HK, HM, and of the target surface were 74.2, 3,775, 2436.6, and 55,000 MPa, respectively. Poisson's ratio of EXO and HK is 0.45 and 0.3.

The plan model was introduced to present the contact mechanics of a grasshopper's attachment pads to the target surface. The target surface was defined as a rigid body. The EXO was defined as two-dimensional hyper-elastic solid (HYPER56). The HK was defined by a two-dimensional, eight-node structured solid (PLANE82). HM was defined as a two-dimensional contained element (Fluid79). The contact was defined by a two-dimensional general contact element (CONTACT – 48).

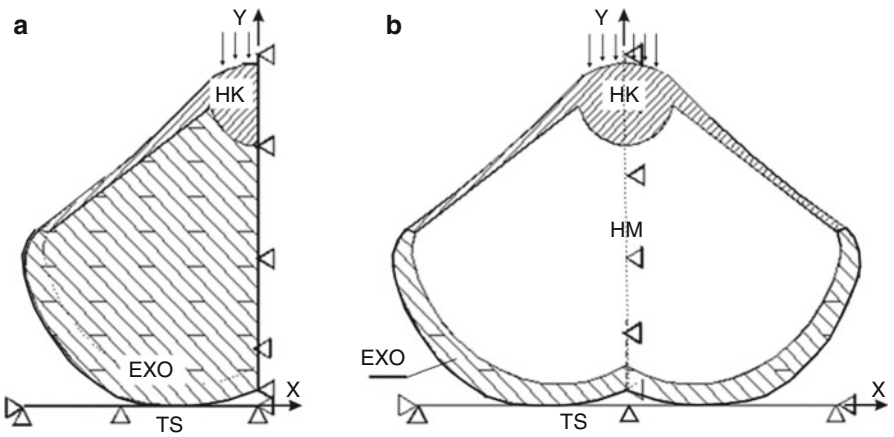
The target surface TS was fully restrained $\Delta X = \Delta Y = 0$. Displacement on the axis of symmetry Y was zero and set at $\Delta X = 0$ because both structure and load were symmetrical. The reaction force is from ten to hundreds of micro-Newtons. The preloading displacements were determined in order to cover the range of jumping force. Here, the loads were set up as restrained displacement on the HK surface from 0.01 to 0.1 mm (see Fig. 9 for more details). The estimation was obtained under the presumption that the jump was performed in micro-seconds.

Figure 10a and b show the deformation vectors of the grasshopper's attachment pad, Fig. 10c and d are detailed images in the contact zone (marked by squares in a and b) for the SS and FC models, respectively. The displacement vector fields were obtained under a restrained displacement of 0.1 mm at restrained HK surface. In order to show the deformation clearly, the displacement of each node points is magnified ten times. For the SS model, the deformation vectors in the contact zone are perpendicular to the target surface (Fig. 10c), which suggests that no relative movement between pad and target surface is generated during the contact process, so it is reasonable to assume that no friction force is generated between the surfaces. On the other hand, the deformation vectors in the contact zone (Fig. 10b, square zone) in the FC model are parallel to the target surface (Fig. 10d) and the directions of displacement vectors in two contact zones (Fig. 10b) are reversed and are symmetrical in the Y axis. So it is reasonable to assume that friction forces in the reversed direction on contacted zones are created during the contact process and the frictional force increases with increasing loads. The generated friction forces, compared with that on the contact zones of SS model, would enhance the stability of the contact.

It was believed that both geometric structure and the material topology make the displacement vector fields so different. The geometric design made it possible for the pads to move outside, but the tendency was more strongly restrained by the soft material (SS model) than that of fluid contained structures (FC model). This behavior results from the mechanical properties of two materials. As a fluid, hemolymph can bear only

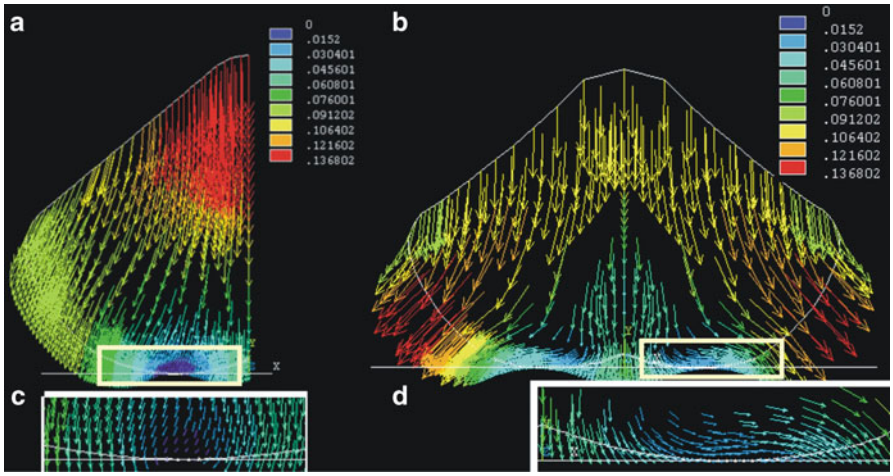


Adhesion in the Animal World, Fig. 8 Grasshopper's attachment pads (**a**) and an illustration of the cross section (**b**) on third tarsmere. (**a**) Tarsus includes tarsomeres (*T1–T3*) with attachment pads and claw (*T4*); *Cs* cross section; (**b**) Diagram of cross-section of third tarsomere. *AS*, airbag; *HM*, hemolymph; *CL*, epidermal cells; *HK*, hard cuticle; *TD*, flexor tendon of claw; *EXO*, exocuticle with rod structure supporting the superficial layer



Adhesion in the Animal World, Fig. 9 Finite element models. (**a**) Simplified model (SS). Only two materials, soft excuticle, *EXO*, and hard cuticle, *HK*, are considered. *TS* target surface. (**b**) Fluid-contained model (FC). Three materials, *EXO*, *HK*, and hemolymph, *HM*, are considered. Fluid material *HM* in contained by *EXO* and *HK*. Target surface is rigid solid and stable-restrained in both X and Y directions. Symmetry axis Y of both models is restrained in X direction

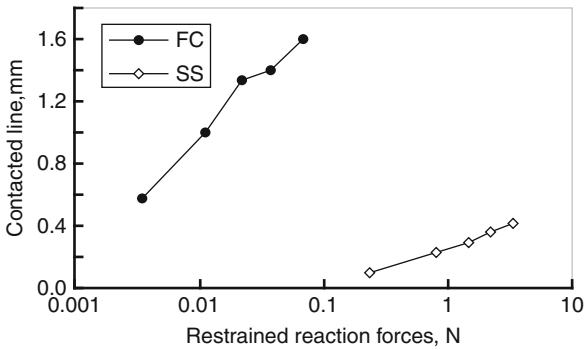
compressive stress, but not the shear stress and tensile stress. The compressive stress generated by the ground reaction force on contact zone and HK boundary was transmitted by hemolymph to the outside of container-like structure, leading to the displacement of node point in FC model in reversed directions (toward outside). On the contrary, the soft material can bear shear, tensile, and compressive stresses, which restrains the motion of the node point in the SS model to move outside. The tendency will result the shear force in reverse directions on the contacted area and the force would make the attachment more stable. On the other hand, the spring constants in the load direction are K_s = 34.44 N/mm and K_f = 0.021 N/mm for the SS model and FC models, respectively. The difference between the two models, K_s/K_f, is 1,640. This means that the geometric design and the material topology of a grasshopper's pad minimizes stiffness and makes the pad much flexible. Lower stiffness



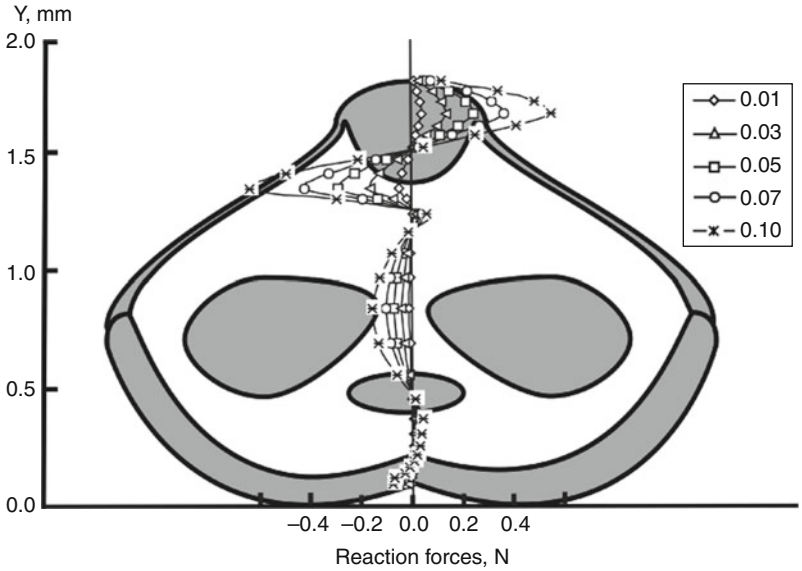
Adhesion in the Animal World, Fig. 10 Vector field of displacements. (**a**) SS model and (**c**) detail of contact zone. The vectors in contact zone are perpendicular to the target surface; there is no relative movement between pad and target surface, so no friction force is generated during contact. (**b**) FC model and (**d**) detail of contact zone. The vectors in the contact zone are parallel to the target surface; a relative movement at reversed directions between pad and target surface is observed and thus reversed friction force will be generated during contact. The displacement vectors near contact zone in FC model (**b** and **d**) are much larger than those in SS model (**a** and **c**), suggesting that the increased rate of contact area in the FC model is larger than in the SS model

would greatly decrease the impact force during landing and prevents other parts of the leg from over being loaded. Large impact force is one of the major causes of failure in transmission gears in legged robots. This suggests that the geometric design of the grasshopper foot may be applicable to the design of legged robots as it may reduce impact forces.

Lower stiffness also means that the pad is more flexible, and larger contact area can be obtained during attachment. Figure 11 shows the contact area for both the SS and FC models, where the lengths of contacted lines were obtained by checking the reaction forces on the nodes of the contact elements. The zones were calculated by the length of contacted line when the reaction force on the node is not zero. Although the reaction force (namely the load acted) of FC model is smaller than that of SS model, the contacted line is much longer than that for the SS model, suggesting that in the fluid contained geometric design (FC model) it may be possible to increase adhesive force on the surface when grasshopper's attachment pads are attached on various surfaces, which is irrelevant to the adhesive mechanisms. A large real contact area always helps to obtain strong adhesive force. This analysis may answer why grasshoppers can make their locomotion on upside-down surfaces. The results suggest that by designing the geometry of the artificial grasshopper tarsal, researchers may not only decrease landing impact force but also increase the contact area and thus increase the adhesion between grasshopper pads and target surface. Both of these characteristics are needed in designing legged robots.



Adhesion in the Animal World, Fig. 11 Relationship of contacted area with pre-restrained distance and restrained reaction forces. Contacted line *Lc* versus restrained reaction forces Frr. Results show that $Lc = 0.3505 \times Ln(\text{Frr}) + 2.5839$, $R^2 = 0.9815$ for FC model and $Lc = 0.118 \times Ln(\text{Frr}) + 0.2619$, $R^2 = 0.991$ for SS model. These results mean that with the same restrained displacement, reaction forces of FC model are much lower than that of SS model, but the contact line are much higher than in SS model



Adhesion in the Animal World, Fig. 12 Reaction forces at restrained points for different preloads. With increasing loads, the restrained reaction forces are also increased. But the forces are always zero at the place where the flexor tendon is located

The reaction forces at restrained points in the *Y* direction (Fig. 12) show that with an increase in load, the reaction force also increases. The biggest reaction force is located in the HK zone and nearby, but in reverse direction, in the rod based tissue exocuticle that supports the superficial layer. The reaction force in the HM zone is lower than in the neighboring exocuticle. Interestingly, the restrained force in the tendon area is zero. This result could explain why the tendon can keep its position in the hemolymph.

The macro-shape of pad and topological distribution of the material have inspired engineers to create feet for legged robots at Nanjing University of Aeronautics and Astronautics for reducing the impact force during attaching and increase the contact stable. Both characters are important for legged horizontal running robot. The wall climbing robot based on wet adhesion was studied but still no robot was present.

Gecko Mimicking Robot: Reducing the Contact Stiffness

To integrate enough adhesive force by van der Waals interaction, it is a key technique to decrease the repulsive force by lowering the contact stiffness. Figure 13 show a model for calculating the contact stiffness of a hair along and perpendicular to the seta when it is vertical (a) or with a slope angle α (b), the curve of stiffness with slope angle was present in Fig. 13c.

The spring constant along the cantilever beam (Fig. 13a), k_a

$$k_a = \frac{\pi}{4}\frac{Ed^2}{l} \tag{4}$$

And that perpendicular to the cantilever beam (Fig. 13a), k_p

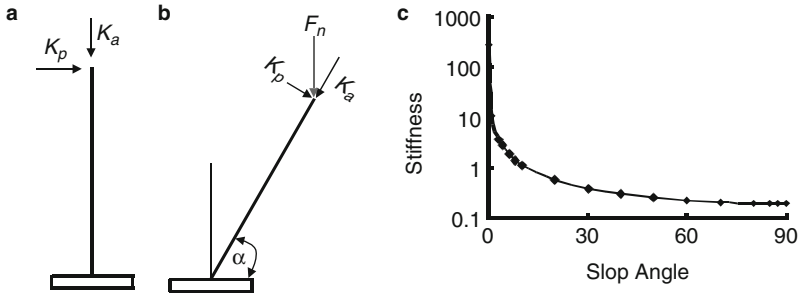
$$k_p = \frac{3\pi Ed^4}{64l^3} = \frac{3\pi}{64}\frac{Ed^2}{l}\left(\frac{d}{l}\right)^2 \tag{5}$$

In equations 4 and 5, E, d, and l are the elastic modulus, the diameter and the length of cantilever beam, respectively. To understand the effects of force direction on the stiffness, the stiffness k_a to k_p was compared. The ratio becomes

$$\frac{k_a}{k_p} = \frac{16}{3}\left(\frac{l}{d}\right)^2 \tag{6}$$

Bringing the typical elastic modulus and the geometric of gecko seta into above equations, Table 1 can be obtained.

The results suggest great differences between contact stiffness along k_a and perpendicular k_p and the gecko seta. The stiffness along the hair is up to 10,000 times the stiffness perpendicular to the hair, which means that when forces act on the direction perpendicular to the hair, the hair would be very soft and have more points coming into contact. The results explain why all animals'



Adhesion in the Animal World, Fig. 13 Contact model and stiffness. (**a**) Mechanical model for a vertical cantilever; (**b**) Mechanical model for a slope cantilever; and (**c**) Contact stiffness versus slope angle of cantilever

Adhesion in the Animal World, Table 1 The stiffness of gecko setae

E (GPa)	d (μm)	l (μm)	K_a (N/m)	K_p (N/m) × 10^{-3}	K_a/K_p
1	1	30	26.18	0.0054	4,800
	5	130	151.04	0.0419	3,605

adhesive hairs are settled in slope. When a hair slopes to the base flat surface on an angle, α, a relationship of the stiffness, k, in the force direction is set up

$$k = \frac{1}{\sqrt{\left(\frac{\sin\alpha}{k_a}\right)^2 + \left(\frac{\cos\alpha}{k_p}\right)^2} \cdot \cos(\alpha - \gamma)} \qquad (7)$$

where $\tan\gamma = \frac{k_p \sin\alpha}{k_a \cos\alpha}$. The relationship is shown in Fig. 13c and suggests that when force acting on a direction shifts away from along the hair's very small angle, the stiffness will decrease dramatically (decrease to 1% when slope angle is only 4°). In nature, the gecko's setae array are sloped to the surface from 27° to 70°, and during contact, the toes slide on the target surface. This procedure further increases the slop angle of gecko hairs and results a decrease of contact stiffness. Dai's group measured the three-dimensional reaction forces of toe/foot of freely moving gecko on floor, wall and ceiling by using a newly developed force measuring array [Dai, et al. 2011]. They found that the preload force acting on a toe would be no more than 50 mN, but the shear force generated during attachment of a toe on a substrate would be several hundreds mN [Wang et al. 2010]. Moreover, the shear force is always along the toe and points from the center of foot to the terminal of toe. On the other hand, gecko's first and fifth toe almost locate in reverse direction, which make the shear force acting on the toes being mechanically redundancy. This design greatly increases the attachment reliability, which are very similar to the soft pads of grasshopper.

Gecko-mimicking dry adhesive has attracted much research in the past 10 years. Artificial setae array based on nano-carbon tube array and polymer were developed; they showed adhesion ten times greater than that of natural gecko seta. Some researchers tried to use artificial adhesive in a gecko mimicking robot, but unfortunately the moving ability of those robots still lags far behind that of a natural gecko.

Tree-Frog Toe Mimicking Tire

The structure and contact mechanics of the foot of tree frogs was studied by Barnes. Inspired by the results and discoveries from a cat's paw, Continental AG has obtained safer braking (Barnes 2007). During driving, the tire receive a widened contact to ensure safe transmission of greater forces to the road. The tire is therefore able to transmit braking forces more efficiently than a conventional tire. Under normal driving conditions, the tire remains slim to provide protection against aquaplaning.

In short, adhesion of animals is nature's excellent design. Animals have integrated fine micro-structures and structures, and are elaborate physical mechanisms. The geometric design and use of the mechanisms may inspire humans to develop more smarter system for a better life.

Cross-References

- ▶ Adhesion Hysteresis
- ▶ Basic Concepts in Adhesion Science
- ▶ Capillary Force and Surface Wettability
- ▶ Gecko Toe Surface
- ▶ Interfacial Energy
- ▶ Polymer Adhesion
- ▶ Tribological Characteristics of Insects' Skin
- ▶ Surface Forces, Surface Tension, and Adhesion

References

E. Arzt, S.N. Gorb, R. Spolenak, From micro to nano contacts in biological attachment devices. Proc. Natl. Acad. Sci. U.S.A. **100**(19), 10603–10606 (2003)
K. Autumn et al., Adhesive force of a single gecko foot-hair. Nature **405**, 681–685 (2000)
K. Autumn et al., Evidence for van der Waals adhesion in gecko setae. Proc. Natl. Acad. Sci. U.S.A. **99**, 12252–12256 (2002)
W.J.P. Barnes, Biomimetic solutions to sticky problems. Science **318**(12), 203–204 (2007)
R.G. Beutel, S.N. Gorb, Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. J. Zool. Syst. Evol. Res. **39**, 177–207 (2001)
C.J. Clemente, W. Federle, Pushing versus pulling: division of labor between tarsal attachment pads in cockroaches. Proc. R. Soc. B **275**, 1329–1336 (2008)
Z.D. Dai, S.N. Gorb, A study on contact mechanics of grass-hopper's pad (Insecta: ORTHOPTERA) by finite element methods. Chinese Sci. Bull. **54**(4), 549–555 (2009)
Z.D. Dai et al., Roughness-dependent friction force of the tarsal claw system in the beetle *Pachnoda marginata* (Coleoptera, Scarabaeidae). J. Exp. Biol. **205**, 2479–2488 (2002)
Z.D. Dai et al., Dynamics of gecko locomotion: a force measuring array to detect 3D reaction forces. J. Exp. Biol. **214**, 703–708 (2011)
M.H. Dickinson et al., How animals move: an integrative view. Science **288**(7), 100–106 (2000)
W. Federle et al., Attachment forces of ants measured with a centrifuge: better 'wax-runners' have a poorer attachment to a smooth surface. J. Exp. Biol. **203**, 505–512 (2000)
W. Federle et al., An integrative study of insect adhesion: mechanics and wet adhesion of pretarsal pads in ants. Integr. Comp. Biol. **42**(6), 1100–1106 (2002)
S.N. Gorb, The design of the fly adhesive pad: distal tenent setae are adapted to the delivery of an adhesive secretion. Proc. R. Soc. Lond. B. **265**, 747–752 (1998)
G. Huber et al., Evidence for capillarity contributions to gecko adhesion from single spatula nanomechanical measurement. Proc. Natl. Acad. Sci. U.S.A. **102**(45), 16293–16296 (2005)
J.N. Israelachvili, *Intermolecular and Surface Forces* (Academic, London, 1985), pp. 90–91
Y.S. Jiao et al., Adhesion measured on the attachment pads of *Tettigonia viridissima* (Orthoptera, Insecta). J. Exp. Biol. **203**, 1887–1895 (2000)
B.N.J. Persson, Wet adhesion with application to tree frog adhesive toe pads and tires. J. Phys. Condens. Matter **19**, 1–6 (2007)
Z.Y. Wang et al., Morphology and reaction force of toe of geckos freely moving on ceiling and walls. Sci. China E. **53**(6), 1688–1693 (2010)

Adhesion in the Contact of Polymers

► Polymer Adhesion

Adhesion of Elastic Spheres

► Adhesive Contact of Elastic Bodies: The JKR Theory

Adhesion of Polymers

► Polymer Adhesion

Adhesion of Spherical Particles

► Adhesive Contact of Elastic Bodies: The JKR Theory

Adhesive Contact

► Adhesive Contact of Elastic Bodies: The JKR Theory

Adhesive Contact of Elastic Bodies

► Adhesive Contact of Elastic Bodies: The JKR Theory

Adhesive Contact of Elastic Bodies: The JKR Theory

K. L. Johnson, J. A. Greenwood
Department of Engineering, University of Cambridge, Cambridge, UK

Synonyms

Adhesion of elastic spheres; Adhesion of spherical particles; Adhesive contact; Adhesive contact of elastic bodies

Definition

In 1882 H. Hertz published his famous paper, *Uber die beruhrung feste elastischer Korper* ("On the contact of elastic bodies") *J. reine und angewandte mathematiik*, **92**, 156–171. The bodies were assumed to have smooth curved surfaces, so that only normal compressive forces are transmitted between them. In this entry, the effect of tensile traction, arising from molecular adhesive forces at the interface is presented. The solution lies in the field of the linear theory of elastic solids.

Key Applications

Applications are to be found in the adhesion of small spherical particles, especially in the food-processing industry and in micro/nano scale tribology.

Rigid Spheres

When two smooth surfaces are placed in close proximity h, they attract each other by a force per unit area $\sigma(h)$. The work done to separate the surfaces is known as the surface energy or work of adhesion $\Delta\gamma = \int_{z_0}^{\infty} \sigma(h)dh$, where z_0 is the equilibrium separation. When the surfaces are those of two rigid spheres of radii R_1, R_2, the separation between the two close to the point of closest approach may be written

$$h = h_0 + r^2/2R \tag{1}$$

where $R = R_1R_2/(R_1 + R_2)$. Derjaguin (1934) showed that the force of adhesion between the spheres will be $P_a = \int_0^{\infty} \sigma(h)\, 2\pi\, r\, dr = 2\pi \int_{h_0}^{\infty} \sigma(h)\, R\, dh$ and, if $h_0$is the equilibrium separation z_0, the force will be $P_a = 2\pi R\Delta\gamma$, independently of the particular law of surface force. This recovers rather readily the result found earlier by Bradley (1932) by integrating the molecular attractions between every pair of molecules over the volumes of the spheres.

Bradley (1931) had previously shown how the interaction energy ϕ between molecules was related to the force between plane surfaces and so to the surface energy: if $\phi = A/r^m$, then the force/unit area is $\sigma(h) = \frac{2\pi q^2 A}{(m-2)(m-3)} \frac{1}{h^{m-3}}$ (where q is the density of attracting molecules in each solid), so that the "6–12" Lennard-Jones interaction potential gives rise to a "3–9" law of force between surfaces.

Elastic Spheres

Strong interest existed in the adhesion of *elastic* spheres, initially between small particles in colloidal solution and more recently with microprobe instruments such as the atomic force microscope. The first analysis was by Derjaguin (1934). For publication in English, see Derjaguin et al. (1975), referred to as the "DMT theory." Surprisingly, this theory gave the force to separate the spheres as $2\pi R\Delta\gamma$: the same as for rigid spheres.

Independently, without knowledge of the Russian work, Johnson et al. (1971) were investigating the adhesion of highly elastic spheres of rubber from a different point of view. Under the action of a compressive force P_h the spheres are compressed to a contact of radius a given by the Hertz theory, at which they adhere. The subsequent application of a tensile load causes the surfaces to peel apart in the manner of an elastic crack. The application of elastic fracture mechanics (see below) showed that the surfaces spring apart at a net force $(3/2)\pi R\Delta\gamma$, which is different from the DMT result. At first these results appeared to be conflicting, but Tabor (1977) realized that they were asymptotic results for rigid spheres on the one hand and highly elastic spheres on the other. He proposed a distinguishing parameter;

$$\mu \equiv \left(\frac{R(\Delta\gamma)^2}{E^{*2}z_o^3}\right)^{1/3} \tag{2}$$

where the combined modulus of the spheres $E^* = \left[(1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2\right]^{-1}$. μ can be shown to express the ratio of the elastic deformation of the spheres at the point of separation to the range of action of adhesive forces. At high values of μ the adhesive forces effectively act at the interface. At values of μ approaching zero, the spheres do not deform significantly under the action of the adhesive forces.

The JKR Theory

In the original publication (Johnson et al. 1971) an expression was derived for the total free energy of the system comprising elastic, surface, and gravitational components. Differentiating this expression gave the contact radius a in equilibrium at any given load P. Subsequently, Maugis and Barquins (1978) showed how the same results could be obtained readily using elastic fracture mechanics, as follows.

In the conceptual model the spheres are pressed into contact by a normal load P_h, which, in the absence of adhesion, results in a contact radius a_h, a pressure distribution $p_h(r)$, and a compression δ_h given by the Hertz theory:

$$\begin{aligned} a_h^3 &= 3RP_h/4E^*; \\ p_h(r) &= (3P_h/2\pi a^2)\{1 - r^2/a^2\}^{1/2}; \\ \delta_h^3 &= (a^2/R)^3 = (9P_h^2/16RE^{*2}) \end{aligned} \tag{3a, b, c}$$

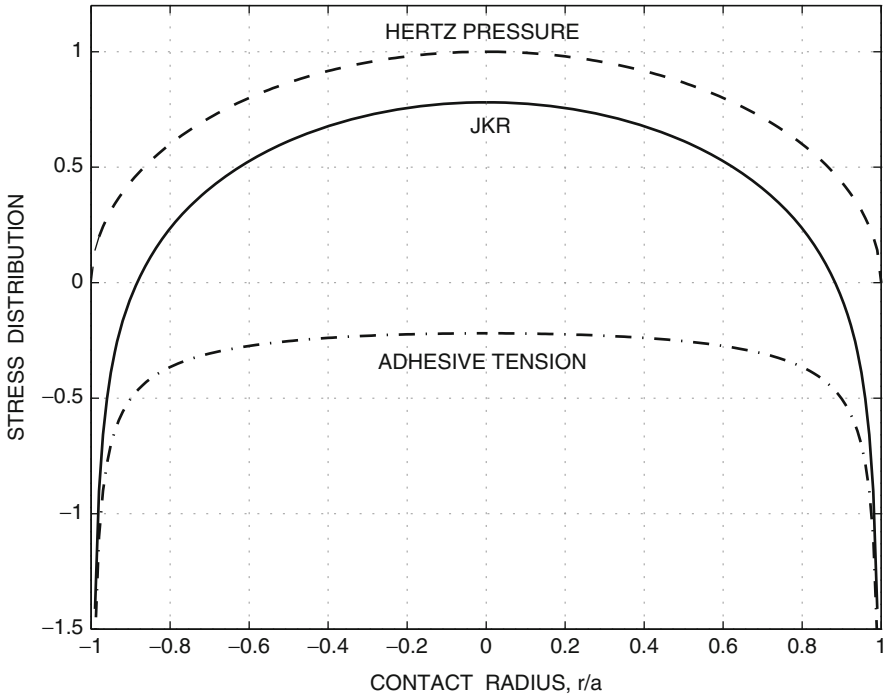
The contact surfaces are then assumed to adhere while a tensile force P_a is applied to the contact, resulting in a Boussinesq distribution of *tensile* (negative) pressure:

$$p_a(r) = \left(P_a/2\pi a^2\right)\left\{1 - r^2/a^2\right\}^{-1/2} \tag{4}$$

The resultant pressure $p(r) = p_h(r) - p_a(r)$ is shown in Fig. 1.

The adhesive term p_a has a singularity at the edge of contact that corresponds to a stress intensity factor $K_I = P_a/2a\sqrt{\pi a}$. Equating the elastic energy release rate G to the work of adhesion $\Delta\gamma$ and using Irwin's relationship: $G = K_I^2/2E^* = \Delta\gamma$, gives $P_a = (8\pi\,\Delta\gamma\, E^* a^3)^{1/2}$so that the net load

$$P = P_h - P_a = 4E^* a^3/3R - \sqrt{8\pi\,\Delta\gamma\, E^* a^3} \tag{5}$$



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 1 JKR pressure distribution comprises a Hertz pressure less a Boussinesq (rigid punch) distribution

It is convenient to introduce non-dimensional variables $P^* = P/\pi R\,\Delta\gamma$; $a^* = a(4E^*/3\pi R^2\Delta\gamma)^{1/3}$; $\delta^* = \delta/\left\{9\pi^2 R\,\Delta\gamma^2/16E^{*2}\right\}^{1/3}$ in normalized form, (5) becomes:

$$P^* = P_h^* - (6P_h^*)^{1/2} = a_h^{*3} - (6a_h^{*3})^{1/2} \qquad (6)$$

It is plotted in Fig. 2 and shows the equilibrium contact size a under a given load.

Point C ($P_c^* = -3/2$, $a_c^* = (3/2)^{1/3}$) is a critical point at which the surfaces snap apart – the pull-off point. In real variables, $P_c = (3/2)\pi R\,\Delta\gamma$, as quoted above.

An alternative procedure is to control the displacement δ, rather than the load P, a procedure known as "fixed grips." The net displacement at a contact radius a may be written $\delta = \delta_h + \delta_a$, where $\delta_h = a^2/R$ is the Hertz compression and $\delta_a = P_a/2\pi E^* = \sqrt{8\pi E^*\Delta\gamma\, a^3}/2\pi E^*$, so that:

$$\delta = a^2/R - \sqrt{2\pi\,\Delta\gamma\, a^3/\pi E^*} \qquad (7)$$

This equation can be normalized using the non-dimensional variables given above to read:

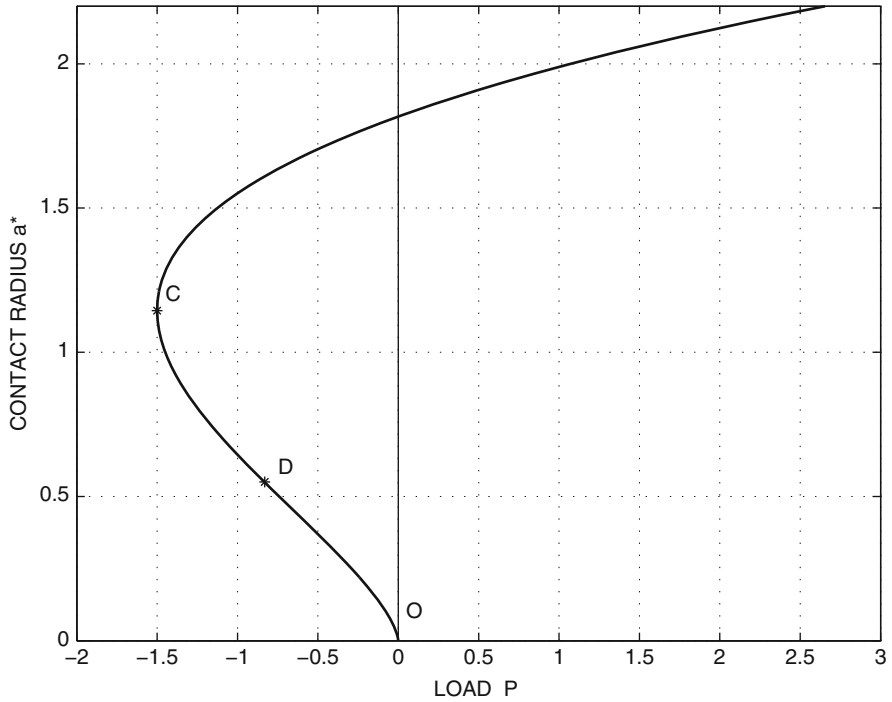
$$\delta^* = a^{*2} - 2\sqrt{2a^*/3} \qquad (8)$$

Equations (6) and (8) are parametric equations for load as a function of displacement as plotted in Fig. 3.

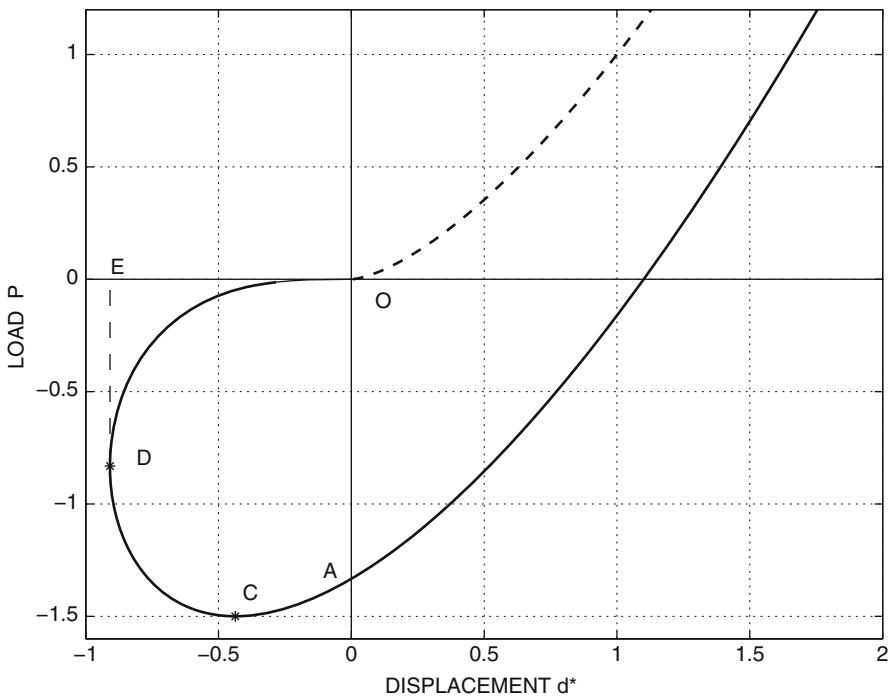
As the surfaces are brought together, the load P^* remains zero until the surfaces touch at $\delta^* = 0$, whereupon they jump instantaneously into contact at point A $(0, -4/3)$. Subsequent positive displacement causes the load to increase positively as shown. If the displacement is now reversed, it decreases reversibly to a minimum at point D $(-0.909, -0.833)$, at which a further jump occurs to point E on the load axis. Both jumps, into and out of contact, involve energy dissipation in the form of high frequency elastic waves.

Perfect fixed grips are difficult to achieve practically. Some degree of compliance is usually present, as illustrated by the spring in Fig. 4a.

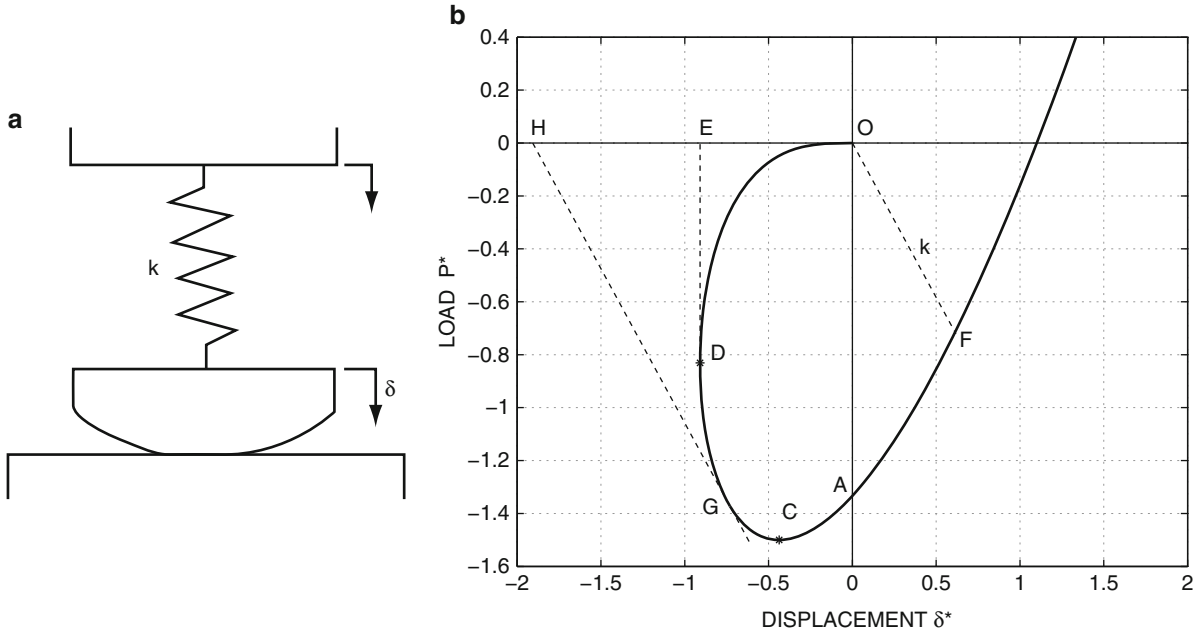
This might be very stiff, arising from the compliance of the grips holding the two spheres, or much smaller, as in the atomic force microscope, where the contacting tips are supported by a flexible cantilever. Assuming the spring in Fig. 4a to be linear and of stiffness k, it exerts an additional force $k\delta$ acting on the grips, which appears in Fig. 4b as a line of gradient k. Starting at the point of first touch, the spheres jump unstably into contact at point F where equilibrium is restored. They then move reversibly with



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 2 Variation of equilibrium contact radius a with load in dimensionless variables. Point C defines the pull-off force in load control, while point D is the unstable point in displacement control



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 3 Variation of equilibrium contact displacement with load, in dimensionless variables. The *broken line* is the load–displacement curve in Hertz adhesionless contact



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 4 Loading-unloading cycle with spring supported grips. (**a**) Grips held by a linear spring. (**b**) Loading: unstable jump into contact OF; unloading: FACG: jump out of contact GH

increasing force. On unloading, the displacement moves stably to point G, which is tangential to the spring stiffness k. The surfaces then jump apart to point H at zero load. This behavior can be expressed in the non-dimensional variables used in Fig. 3 as follows:

$$P_s = k\delta$$

where P_s is the compressive force exerted by the spring, so that:

$$P_s^* = \delta^* \{k[(16/9)\pi R^2 E^{*2} \Delta\gamma]^{-1/3}\} \tag{9}$$

where the curly bracket defines a non-dimensional spring stiffness.

The Transition from Rigid (*Bradley*) to Elastic (*JKR*)

The transition in the value of the force to separate the spheres (pull-off force) from $2\pi R\,\Delta\gamma$ for rigid spheres to $(3/2)\pi R\,\Delta\gamma$ for highly elastic spheres was explored numerically by Muller et al. (1980) and Greenwood (1997). The infinite adhesive stress at the periphery of the contact in the JKR model must be relieved in practice by a slight separation of the surfaces. This effect was modeled in closed form by analogy with a "Dugdale crack" by Maugis (1992). The adhesive stress distribution is illustrated in Fig. 5. The surfaces separate in the Dugdale zone $d = c - a$, where the adhesive stress is assumed to be constant σ_o. Thus, the work of adhesion $\Delta\gamma = \sigma_o h_o$. Maugis introduces a transition parameter $\lambda = \left(9R\sigma_o^3/2\pi E^{*2}\Delta\gamma\right)^{1/3}$. The maximum stress $\sigma_o$was taken to be the maximum of the Lennard-Jones stress, as shown in Fig. 5, which gives $h_o = 0.97 z_o$ and (by (1)) $\lambda = 1.16\mu$.

Maugis obtains the following equation for inner and outer contact radii a and c in terms of λ:

$$(\lambda a^{*2}/2)\left\{(m^2-2)\mathrm{s}^{-1}m + \sqrt{m^2-1}\right\} + (4\lambda^2 a^*/3) \times \left\{\sqrt{m^2-1}\,\mathrm{s}^{-1}m - m + 1\right\} = 1 \tag{10}$$

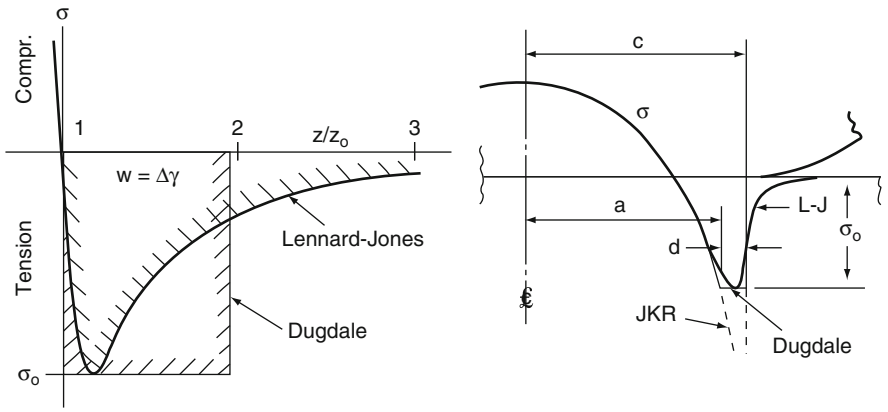
where m = c/a. For the load $P^* = P_h^* + P_a^*$:

$$P^* = a^{*3} - \lambda a^{*2}\left\{\sqrt{m^2-1}\right\} + m^2\mathrm{s}^{-1}m \tag{11}$$

The variations in a^*, c^*, and P^* with λ are found by the simultaneous solution of (10) and (11), and for the displacement (see Fig. 6):

$$\delta^* = a^{*2} - (4/3)\lambda\sqrt{m^2-1} \tag{12}$$

The pull-off force in load control is found from the minimum (negative) force. This graph is added to Fig. 2, where it compares favorably with the computed results based on Lennard-Jones. At values ofλ less than ≈0.1 Maugis' result approaches the rigid asymptote $2\pi R\,\Delta\gamma$;

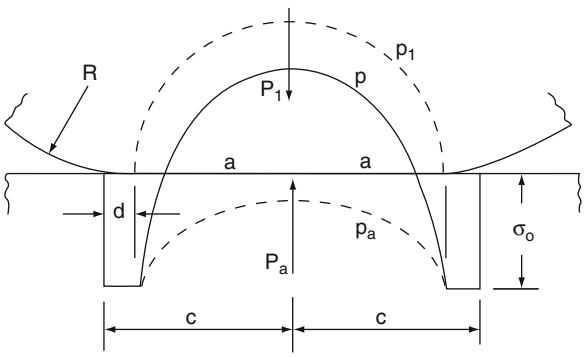


Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 5 Maugis–Dugdale model. The surfaces separate slightly at the edge of contact where the adhesive stress is assumed to be constant

at values of λ greater than ≈5, it approaches the JKR result $(3/2)\pi R\Delta\gamma$. The difficulty in applying the Maugis–Dugdale model to experimental results is due to the absence of an explicit expression for the radius a in terms of the load P. A procedure for overcoming this drawback is presented by Carpick et al. (1999) (see Appendix below).

It must be kept in mind that both the rigid and elastic models involve the Hertz restriction: $a \ll R$.

The unstable jumping into and out of contact for a JKR contact was discussed in the last section. To extend this analysis to the complete range of the transition from JKR to rigid contacts is not so easy, since the Maugis–Dugdale approximation cannot be used. Numerical computations have been performed by Greenwood (1997) and Feng (2000) that show that the magnitude of the jumps decrease as the parameter μ decreases and at $\mu \approx 1$ they disappear (see Fig. 7).



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 6 Maugis–Dugdale surface stress distribution. Note that the assumed stress in the Dugdale zone (c–a) is the maximum σ_o of the Lennard-Jones distribution

Adhesion of Elastic Cylinders in JKR Line Contact

This is the case of highly elastic bodies where the value of λ in the Maugis plot exceeds 5. The analysis is similar to that for the point contact case treated above. In contact along a strip of width $2b$, the net load *per unit length* P is made up of the Hertz load in line contact P_h in equilibrium with the adhesive tensile force P_a. The pressure distribution at the contact can be expressed as:

$$p(x) = \{(\pi/4)E^* b^2/R\}\sqrt{1 - x^2/b^2} - (P_a \pi b)/\sqrt{1 - b^2/a^2} \tag{13}$$

P_a is found as before from the stress intensity factor at the edge of the contact: $K_I = P_a/\sqrt{\pi a}$ and $K_I^2/2E^* = \Delta\gamma$, giving:

$$P = P_h - P_a = \{(\pi/4)E^* b^2/R\} - \sqrt{2\pi E^* \Delta\gamma\, b} \tag{14}$$

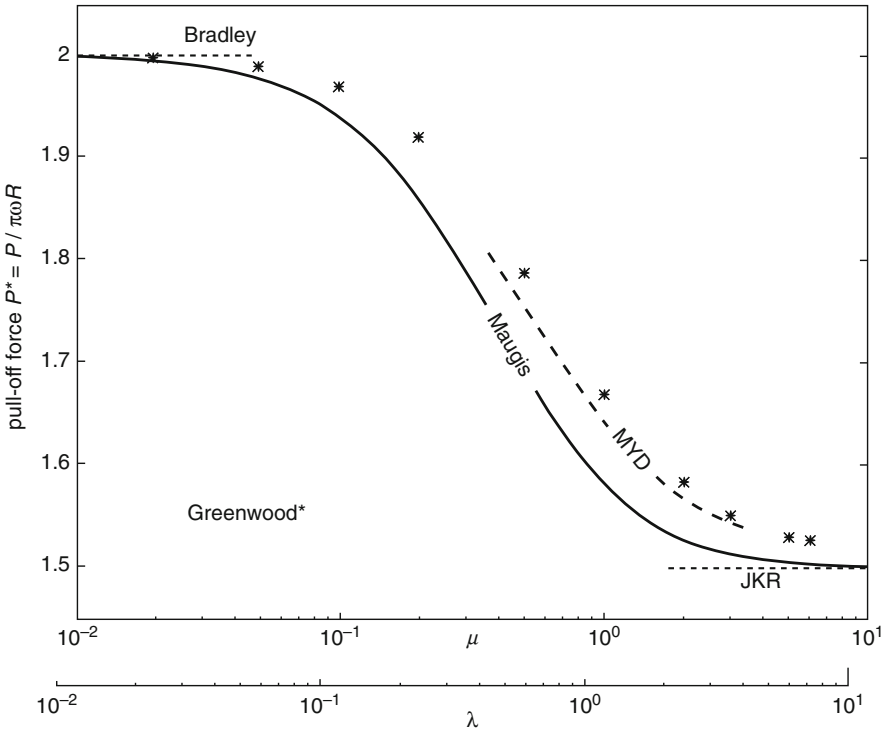
This equation, relating the contact width $2b$ *to* the load P, is the line contact equivalent to (5) in point contact.

Now define a non-dimensional load $P^* \equiv P/\sqrt{2\pi E^* \Delta\gamma}$ and a parameter $K \equiv \sqrt{2R^2\Delta\gamma/\pi E^*}$, so that (1) becomes:

$$P^* = b^2/4K - b^{1/2} \tag{15}$$

The pull-off force P_c^* occurs at the minimum negative load, i.e., when $b = K^{2/3}$ and has the value in real variables:

$$P_c = -(3/4)(4\pi R E^* \Delta\gamma^2)^{1/3} \tag{16}$$



Adhesive Contact of Elastic Bodies: The JKR Theory, Fig. 7 The transition of the pull-off force from rigid (Bradley) to elastic (JKR) materials, as a function of Tabor's parameter μ or Maugis' parameter $.a/(a_o)_\alpha\mu$

Note that, unlike the result for a spherical contact, here the pull-off force does depend on the elastic modulus.

For a Maugis analysis of the transition from rigid to elastic cylinders in line contact, see Johnson and Greenwood (2008).

Appendix

Deduction of Maugis–Dugdale λ from Measurements of Contact Radius

The transition in pull-off force $(P_c)_\lambda$ is shown in Fig. 7. It varies from the DMT to the JKR limit, i.e.,

$$(P_c)_{DMT} = -2\pi R\,\Delta\gamma; \quad (P_c)_{JKR} = -(3/2)\pi R\,\Delta\gamma \tag{17a, b}$$

The contact radius at zero load a_o given by the two theories is:

$$(a_o)_{DMT} = \left(\frac{3\pi R^2\Delta\gamma}{2E^*}\right)^{1/3}; \quad (a_o)_{JKR} = \left(\frac{9\pi R^2\Delta\gamma}{2E^*}\right)^{1/3} \tag{18a, b}$$

From these the variation of contact radius with load is given by:

$$\frac{a}{(a_c)_{DMT}} = (1 - P/(P_c)_{DMT})^{1/3};$$
$$\frac{a}{(a_o)_{JKR}} = \left(\frac{1+\sqrt{1-P/(P_c)_{JKR}}}{2}\right)^{2/3} \tag{19a, b}$$

Equations (19a, b) can be combined to give:

$$\frac{a}{(a_o)_\alpha} = \left(\frac{\alpha+\sqrt{1-P/(P_c)_\alpha}}{1+\alpha}\right)^{2/3} \tag{20}$$

where $\alpha = 0$ corresponds to the DMT limit and $\alpha = 1$ corresponds to JKR limit. (Equation 20) is referred to as the "COS" equation (Carpick et al. 1999). By evaluating the Maugis–Dugdale theory it was shown that, in the range $0.1 < \lambda < 5$, λ is related to α by:

$$\lambda \cong -0.924.\ln(1 - 1.02\alpha) \tag{21}$$

By this equation corresponding values of λ can be found from measured values of $a/(a_o)_\alpha$. In cases where the contact size cannot be measured directly, e.g., the atomic force microscope, λ can be deduced from friction measurements by assuming the friction force to be proportional to the contact area πa^2.

Cross-References

- ► Adhesion Hysteresis
- ► Adhesive Contact of Inelastic Bodies
- ► Contacts Considering Adhesion
- ► Microtribodynamics of Magnetic Storage Hard Disk Drives
- ► Surface Energy and Adhesion
- ► Surface Forces
- ► Van der Waals Forces

References

R.S. Bradley, The moleculat theory of surface energy. Philos. Mag. **11**, 846 (1931)

R.S. Bradley, The cohesive force between solid surfaces and the surface energy of solids. Philos. Mag. **13**, 853 (1932)

R.W. Carpick, D.F. Ogletree, M. Salmeron, A general equation for fittingcontact area and friction Vs load measurements. J Colloid Interface Sci. **211**, 395 (1999)

B.V. Derjaguin, Theorie des Anhaftens kleiner Teilchen. Kolloid Zeitschrift **69**, 155 (1934)

B.V. Derjaguin, V.M. Muller, Yu.P. Toporov, Effect of contact deformations on the adhesion of particles. J. Colloid Interface Sci. **53**, 314 (1975)

J.Q. Feng, Contact behaviour of spherical elastic particles: a computational study of particle adhesion and deformations. Colloids Surf. **172**, 175–198 (2000)

J.A. Greenwood, Adhesion of elastic spheres. Proc. R. Soc. Lond. **A453**, 1277 (1997)

K.L. Johnson, J.A. Greenwood, A Maugis analysis of adhesive line contact. J. Phys. D Appl. Phys. **41**, 155315 (2008)

K.L. Johnson, K. Kendall, A.D. Roberts, Surface energy and the contact of elastic solids. Proc. R. Soc. Lond. **A324**, 30 (1971)

D. Maugis, Adhesion of spheres: the JKR–DMT transition using a Dugdale model. J. Colloid Interface Sci. **150**, 243 (1992)

D. Maugis, M. Barquins, Fracture mechanics and the adherence of viscoelastic bodies. J. Phys. D Appl. Phys. **11**, 1989 (1978)

V.M. Muller, V.S. Yuschenko, B.V. Derjaguin, On the influence of molecular forces on the deformation of an elastic sphere and its sticking to a rigid plane. J. Colloid Interface Sci. **77**, 91 (1980)

D. Tabor, Surface forces and surface interactions. J. Colloid Interface Sci. **58**, 2 (1977)

Adhesive Contact of Inelastic Bodies

J. A. Greenwood, K. L. Johnson
Department of Engineering, University of Cambridge, Cambridge, UK

Synonyms

Plastic contact

Inelastic Contact: Definition

It is rare for an unloading curve to follow the same track as a loading curve. The norm is a hysteresis loop, indicating energy dissipation in the cycle. Unfortunately, there are many features of material behavior that can produce such a response, and few of them can be analyzed theoretically. When adhesion is acting, perhaps the simplest mechanism is that the surface energy gained in making contact may not equal the surface energy needed to break contact. Surfaces react while in contact. Bonds form and molecules entangle, so that in a very real sense, the surfaces that separate are not the surfaces that mated. In an experimental Johnson-Kendall-Roberts (JKR) contact, the loading and unloading curves may both follow the elastic JKR equation, but using different "making" and "breaking" surface energies, essentially independent of speed, but dependent on the dwell time before unloading.

There is also a mechanical way in which the surface may change on contact. The initial surface roughness may have been removed (or occasionally, with strong adhesion, increased!). And it should always be remembered that, except perhaps with gold, metals are protected by oxide layers and adsorbed oxygen or other films, which may be disrupted on contact.

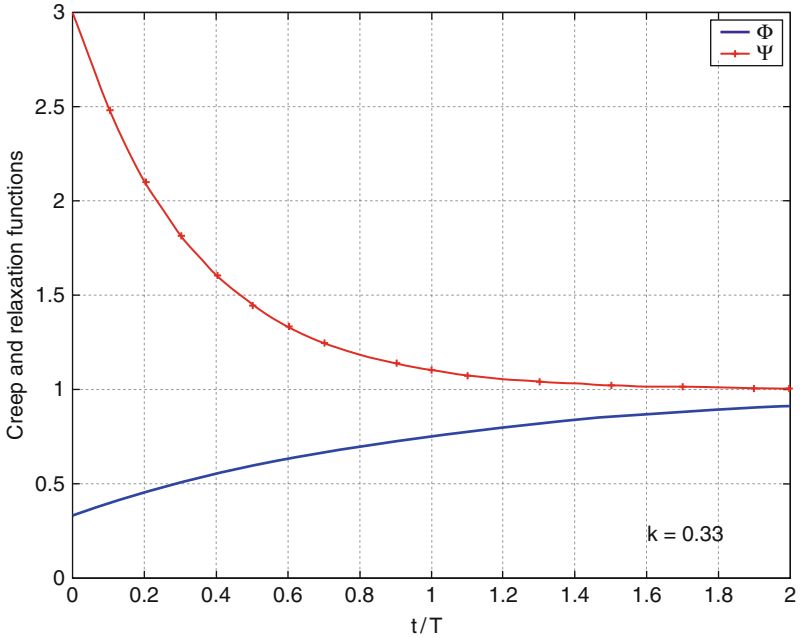
In addition, the material itself may be inelastic. There are many forms of inelastic behavior in solids, generally not amenable to mathematical analysis. The two forms considered here are plastic (or elasto-plastic) behavior, usually associated with metals, where the deformation, or most of it, remains on unloading; and viscoelastic behavior, usually associated with polymers, where the deformation is time-dependent, and will usually disappear after a sufficiently long time.

Linear Viscoelasticity

Scientific Fundamentals

In polymers, behavior is usually time dependent and nonlinear. Linear viscoelasticity, described here, is an idealization showing some of the features of the behavior of real solids; it should never be taken too seriously.

It is assumed that the application of a stress $\sigma(0)$ to the solid produces a strain $\varepsilon(t)$ consisting of an immediate elastic response followed by a gradual increase with time ("elastic creep"), usually settling down to a final equilibrium value. The behavior is described by a creep function $\Phi(t)$ giving the response to a unit stress increment. The simplest forms are the Maxwell solid, which can be regarded as a spring in series with a dashpot, leading to $\Phi(t) = (1/E_0)\{1 + t/T\}$, where E_0 is the instantaneous elastic modulus, and the "three-element solid," a spring in



Adhesive Contact of Inelastic Bodies, Fig. 1 Creep and stress relaxation functions for a three-element solid. Note the differing time constants of creep and stress relaxation

parallel with a dashpot, the pair in series with a second spring. Here, $\Phi(t) = (1/E_\infty)\{1 - (1-k)\exp(-t/T)\}$, where E_∞ is the relaxed modulus. The immediate response is now $\Phi(0) = (k/E_\infty) \equiv 1/E_0$, so that $k = E_\infty/E_0$.

The response to a unit strain increment will also be needed. This is the relaxation modulus $\Psi(t)$ where $\sigma(t) = \varepsilon(0)\,\Psi(t)$. From the representation as springs and dashpots it is easily shown that, for the Maxwell solid, $\Psi(t) = E_0\,\{1 - \exp(-t/T)\}$, while for the three-element solid $\Psi(t) = E_\infty\,\{1 + \left(\frac{1-k}{k}\right)\exp(-t/k\,T)\}$.

(More generally, if the Laplace transforms of the creep and relaxation functions are $\bar{\Phi}(p)$ and $\bar{\Psi}(p)$, then $p\,\bar{\Phi}(p)$ is the reciprocal of $p\,\bar{\Psi}(p)$.)

The ratio k of the two moduli is frequently very small, and since this is also the ratio of the two time constants, it follows that the relaxation response to a strain increment is much faster than the creep response to a stress increment. An example is shown in Fig. 1.

(Experimentally, it is more convenient to measure the response to sinusoidal excitation. Then $\sigma(t) = [E'(\omega) + iE''(\omega)]\;\varepsilon(t)]$. E' and E'' (or more usually their equivalents in shear) are called the "storage" and "loss" moduli. For the three-element solid they are $E' = E_\infty\,\frac{1+k\omega^2T^2}{1+k^2\omega^2T^2}$ and $E'' = E_\infty\,\frac{(1-k)\,\omega\,T}{1+k^2\omega^2T^2}$.)

The importance of *linear* viscoelasticity is twofold. Firstly, superposition can be used to find the response to a sequence of stress increments or to a sequence of strain increments. Writing $\sigma(t) = \sigma(0) + \int_0^t \frac{d\sigma}{dt'}dt'$ and adding the responses to each term:

$$\varepsilon(t) = \sigma(0)\,\Phi(t) + \int_0^t \frac{d\sigma}{dt'}\,\Phi(t-t')\,dt'. \qquad (1a)$$

Similarly

$$\sigma(t) = \varepsilon(0)\,\Psi(t) + \int_0^t \frac{d\varepsilon}{dt'}\,\Psi(t-t')\,dt' \qquad (1b)$$

Importantly, these two relations form a transform pair, so that for *any* pair of functions: if $f(t) = g(0)\Phi(t) + \int_0^t \frac{dg}{dt'}\,\Phi(t-t')\,dt'$ then $g(t) = f(0)\Psi(t) + \int_0^t \frac{df}{dt'}\,\Psi(t-t')\,dt'$

Secondly, the initial response satisfies the usual elastic stress–strain relations, so that an elastic solution to a problem may be converted into a viscoelastic solution simply by substituting it into the appropriate hereditary integral (1).

- Of course, there are two elastic constants, not one. In practice, however, for the usual rubbery materials that exhibit an approximation to linear elasticity, Poisson's ratio is near 0.5 and may be treated as a constant. In particular, in problems of contact with a half-space under normal tractions alone, the only elastic constant appearing is the plane strain modulus $E' \equiv E/(1-\nu^2)$

(or $E' \equiv 2G/(1-\nu) \approx 4G$) and the creep and relaxation functions may be taken as modifying E' rather than E. (Unfortunately, the same notation E' is used for two very different quantities, the *viscoelastic* storage modulus and the *elastic* plane strain modulus.)

Non-adhesive Contact Between a Flat-Ended Circular Punch and a Viscoelastic Half-Space

The simplest indentation problem is that where a rigid, flat-ended punch indents a viscoelastic half-space. A displacement Δ is applied over the region $r < a$ of the surface of a half-space at $t = 0$; at time $t = \tau$ the punch is removed to leave the surface stress-free.

The elastic solution is well known: a pressure distribution $p = K\,(a^2 - r^2)^{-1/2}$ corresponds to a load $P = 2\pi\, aK$ and a displacement $\Delta = \pi\, K/E'$. Then, while the displacement Δ is imposed, the viscoelastic stresses will be $p(t) = (E'\Delta/\pi)\Psi(t)\,(a^2 - r^2)^{-1/2}$ and the load $P(t) = 2\,aE'\Delta\Psi(t)$.

If at time $t = \tau$ the displacement were removed, the load would become $P(t) = 2\,aE'\Delta[\Psi(t) - \Psi(t-\tau)]$ with corresponding stresses on what is now a free surface. Of course, what happens physically when the punch is removed is that the surface does not follow the punch upwards but becomes stress-free. In calculating the behavior, the *spatial* distribution $\sigma(r) = K/\sqrt{a^2 - r^2}$ need not be considered. The relations between the stresses $\sigma_{ij}(r, z)$ and the strains $\varepsilon_{ij}\ (r, z)$ all follow the same pattern, and the remaining quantities follow by integration.

If the displacement were maintained, the stresses would be $\sigma(t) = \varepsilon_0\Psi(t)$, so additional stresses $\sigma(t) = -\varepsilon_0\Psi(t)$ must be applied for all $t > \tau$. Writing the stress as $\sigma(t) = \sigma(\tau) + \int_\tau^t \frac{d\sigma}{d\eta}d\eta$, the resulting strains will be $\varepsilon(t) = \sigma(\tau)\,\Phi(t-\tau) + \int_\tau^t \frac{\partial\sigma}{\partial\eta}\,\Phi(t-\eta)\,d\eta$, or, more conveniently (integrating by parts),

$$\varepsilon(t) = \sigma(t)\,\Phi(0) - \int_\tau^t \sigma(\eta)\frac{\partial}{\partial\eta}\Phi(t-\eta)\,d\eta$$

Substituting $\sigma(t) = -\varepsilon_0\Psi(t)$ and adding these to the original ε_0 gives the strains for $t \geq \tau$ as $\varepsilon(t) = \varepsilon_0\, G(t, \tau)$, where

$$G(t,\tau) = 1 - \Phi(0)\,\Psi(t) + \int_\tau^t \Psi(\eta)\frac{\partial}{\partial\eta}\Phi(t-\eta)\,d\eta, \quad (2a)$$

For the three-element solid, $G(t, \tau)$ is readily evaluated analytically:

$$G(t,\tau) = (1-k)[1 - \exp\{-\tau/kT\}]\cdot\exp\{-(t-\tau)/T\} \quad (2b)$$

Note that the first term relates to the duration of the imposed strain and contains the characteristic time constant for relaxation $T_2 = kT$, while the final exponential term relates to the creep after the strain is no longer imposed and involves the creep time constant. Note also that at $t = \tau_+$, the strain will be $\varepsilon_0 G(\tau, \tau) = \varepsilon_0(1-k)[1 - \exp\{-\tau/kT\}]$, while at $t = \tau_-$, when the strain is still imposed, it will be ε_0. The difference $k\cdot\varepsilon_0\left[1 + \frac{1-k}{k}\exp(-\tau/kT)\right]$ is the immediate elastic response to the removal of the stress $\sigma(\tau) = \varepsilon_0\Psi(\tau)$. (It appears that this simple product form only occurs for the exponential decay of the three-element solid.)

Contact Between a Rigid Sphere and a Viscoelastic Half-Space: No Adhesion

In what follows, it will usually be convenient to include the modulus E'_∞ explicitly and take the creep and relaxation functions to give the dimensionless time dependence, so that now, for example, for the three-element solid $\Phi(t) = \{1 - (1-k)\exp(-t/T)\}$.

According to Hertz, the relation between the contact radius a and the load P is $a^3 = \frac{3}{4}\frac{PR}{E'}$, and the displacements within the contact area are $w(r) = a^2/R - r^2/2R$. If a Hertzian displacement field is imposed at a time t', the immediate response is the Hertzian pressure distribution $(2E'_0/\pi R)\sqrt{a^2 - r^2}$, giving a load $P(t') = \frac{4}{3}\frac{E'_0}{R}a(t')^3$. Subsequently, the load will be $P(t) = \frac{4}{3}\frac{E'_\infty}{R}a(t')^3\Psi(t-t')$, gradually decaying to the final $P(\infty) = \frac{4}{3}\frac{E'_\infty}{R}a(t')^3$.

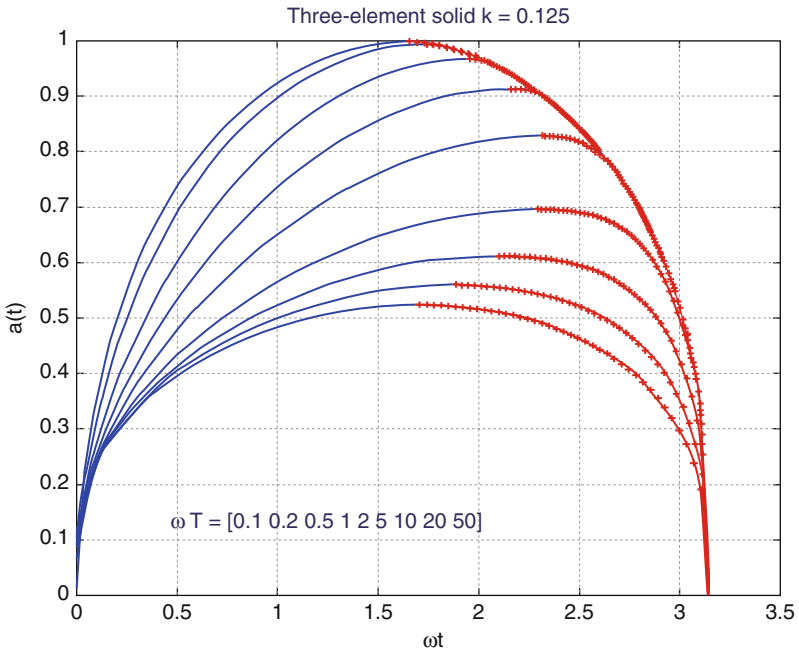
If instead this displacement field is built up from zero, as $a(t)^3 = \int_0^t \frac{da(t')^3}{dt'}dt'$, the result will be

$$P(t) = \frac{4}{3}\frac{E'_\infty}{R}\int_0^t \frac{da(t')^3}{dt'}\{\Psi(t-t')\}\,dt'. \quad (3)$$

Because it is strains (shapes) that are being imposed, the Hertz relation $\delta(t') = a(t')^2/R$ holds at each time, so this equation provides the solution when the approach $\delta(t)$ is specified.

It is important to recognize the implication of (3). Each additional increment of radius implies a corresponding increment of pressure, extending over the contact area $r \leq a$. These pressures decay, but only to values corresponding to the relaxed modulus – they never vanish. The cumulative effect of these decaying pressures represents the actual pressure, *as long as the contact radius exceeds the value it had when the pressures were generated.* But, if the contact radius is less than this value, these decaying *but non-zero* pressures exist *outside* the contact area. The boundary conditions of the problem have then been violated.

To remove them, "negative shapes" must be imposed. But even when the stresses have a simple exponential decay, as in the three-element solid, because of the immediate elastic response this cannot be achieved by applying



Adhesive Contact of Inelastic Bodies, Fig. 2 Contact radius variation during sinusoidal loading. Three-element solid with k = 1/8

a suitably scaled-down shape of radius $a(t)$ at time t. The desired decay rate can be obtained, however, by imposing the negative shape at the instant $t_1$when these stresses were originally stimulated: when the radius $a(t_1)$ during the period when the radius was increasing first reached the current value $a(t)$. In other words, the behavior while the radius exceeded $a(t) = a(t_1)$ should simply be ignored! The stresses, and the load, when the contact radius is $a(t)$ depend purely on the time variation of contact radii up to the time t_1. Thus, (3) is replaced by

$$P(t) = \frac{4}{3}\frac{E'_\infty}{R}\int_0^{t_1} \frac{da(t')^3}{dt'}\{\Psi(t-t')\}\,dt' \qquad (4)$$

where only the upper limit has been changed.

More usually, the load variation is prescribed, so when the load is increasing it is necessary to make use of the transform pair and invert (2) to give

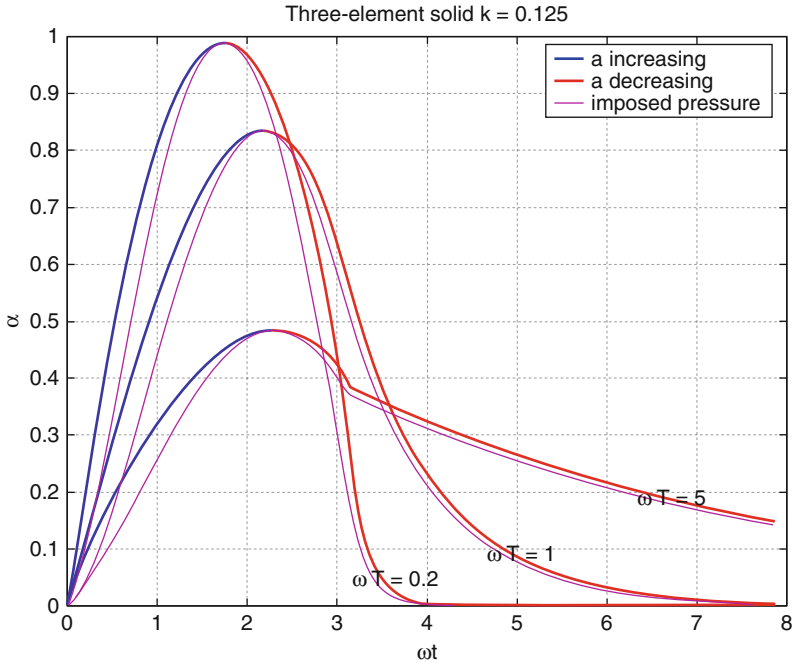
$$a(t)^3 = \frac{3}{4}\frac{R}{E'_\infty}\int_0^t \frac{dP(t')}{dt'}\,\Phi\,(t-t')\,dt' \qquad (5)$$

For a time t during unloading, this equation is used to find the radius variation up to the time t_1. This variation is then substituted into (4) to obtain an equation for the load involving both t and t_1. Equating this expression to the prescribed load $P(t)$ gives an equation from which t_1 can be found. For a detailed example, consult Ting (1966) or Johnson (1985). Figure 2 shows the response of a three-element solid to a sinusoidal load variation $P = P_0 \sin\omega t$ for $0 < \omega t < \pi$. For slow loading, the answer is close to Hertzian behavior with the relaxed modulus E'_∞. For fast loading, it is close to Hertzian with the instantaneous E'_0. In both cases, the behavior is symmetrical between loading and unloading. For intermediate loading rates this is far from the case, and there will be a substantial energy loss in the cycle.

Approach $\delta(t)$

The calculation of the energy loss requires knowledge of the approach $\delta(t)$, and this is also the most readily measured experimental variable. How is it to be calculated? There is no problem during loading. The instantaneous shapes imposed at each instant obey the Hertz relation $\delta(t) = a^2(t)/R$, so are known once the variation of the contact radius is known. But if the same assumption is made as when calculating the load, that during unloading $a(t)$ is replaced by $a(t_1)$, the absurd result is reached that the approach returns to zero when the load and contact radius do. Instead, the pressure distributions at times greater than the time t_m of the greatest contact radius must be calculated from the variation of $a(t_1(\tau))$ using the convolution integral with the stress-relaxation function $\Psi(t-t_1)$.

These stresses must be removed for all times when the contact radius falls below $a(\tau)$, which requires a second



Adhesive Contact of Inelastic Bodies, Fig. 3 Approach resulting from sinusoidal loading $P = P_0 \sin \omega t, 0 < \omega t < \pi$. Also shown is the approach when pressures varying the same way are imposed over a fixed area (scaled to the same maximum value)

convolution integral using the creep function. Using the fact that only the values of the approach during loading occur in the integration, and so taking the Hertzian values $\delta(t_1) = a^2(t_1)/R$, Ting (1966) arrives at

$$R\delta(t) = a^2(t) - \int_{t_m}^{t} \Phi(t-\tau) \left\{ \frac{\partial}{\partial \tau} \int_{t_1(\tau)}^{\tau} \Psi(\tau-\eta) \frac{\partial}{\partial \eta} a^2(\eta)\, d\eta \right\} d\tau \tag{6}$$

It is not clear that (6) has ever been evaluated, and indeed the combination of two integrations with two differentiations is unsuitable for numerical work. Fortunately, Ting's analysis suggests the procedure of rigid punch decomposition: suitable *punch* displacements over $r < a(t_1)$ are imposed at time t_1 and the resulting stresses removed for all times $t > \tau(t_1)$ (recall that t_1 is the time when $a(\tau)$ first occurred). Then, applying the punch solution above (2) to the indentation by a sphere gives

$$R\delta(t) = a^2(t_1) + \int_{t_1}^{t_m} \frac{\partial\, a^2}{\partial\, t'} G(t-t', \tau - t')\, dt'$$

or, for the three-element solid $R\delta(t) = a^2(t_1) + (1-k)\int_{t_1}^{t_m} \frac{\partial a^2}{\partial t'} [1 - \exp\{-(\tau(t') - t')/kT\}] \cdot \exp\{-(t - \tau(t'))/T\}\, dt'$. After the load has returned to zero, this reduces to

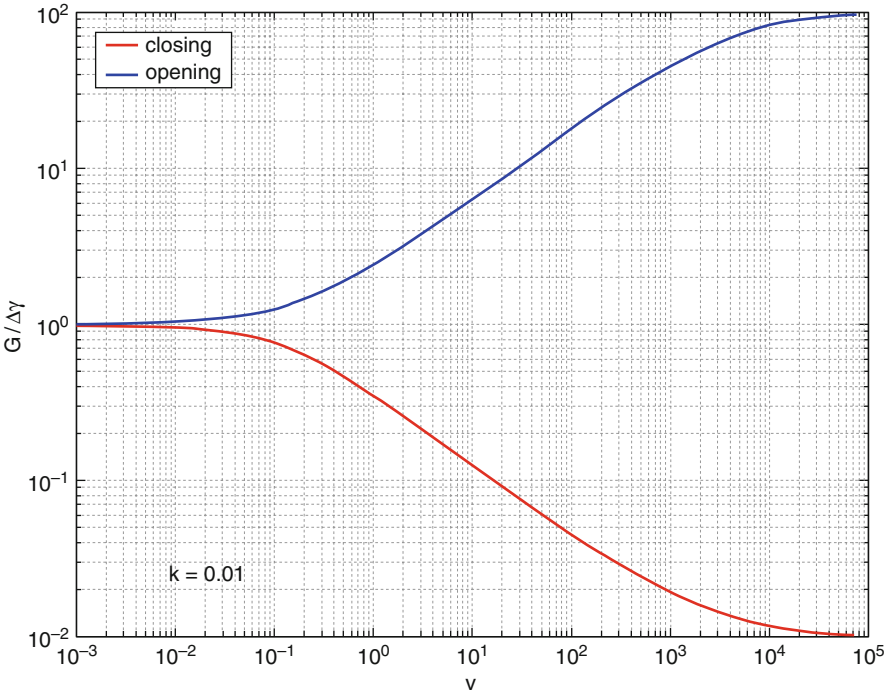
$$R\delta(t) = (1-k)\int_{0}^{t_m} \frac{\partial\, a^2}{\partial\, t'} [1 - \exp\{-(\tau(t') - t')/kT\}] \cdot \exp\{-(t - \tau(t'))/T\}\, dt'$$

These are much more convenient for evaluation than the original solution (6) by Ting, but of course depend on the principle he discovered. Figure 3 shows some typical results. For comparison, the elementary case of displacement due to a sinusoidally varying imposed pressure distribution is shown, revealing that all of the above mathematical complications add rather little!

Adhesive Contact

It may already be clear that extending the above analysis to include the effects of surface energy will not be easy, but a further difficulty arises. The singular stresses at the contact edge give infinite strains, so when the contact edge moves as the contact radius changes, the infinite strain region moves, *giving an infinite strain rate.* The material therefore responds with the *instantaneous* modulus, which is often so large that the contact edge is, in effect, pinned, whatever the loading or unloading rate. This does not correspond to experimental observations.

In order to describe real behavior, the traditional form of fracture mechanics with the stress intensity factor



Adhesive Contact of Inelastic Bodies, Fig. 4 Relation between apparent surface energy and crack speed

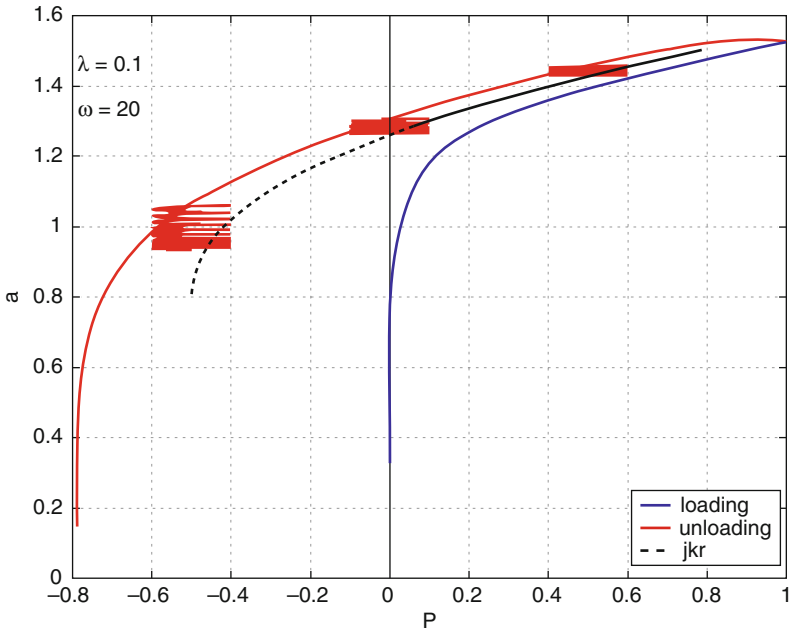
regarded as actually existing must be replaced by the Barenblatt concept, in which the stress intensity factor describes the effects of external loading, but must be cancelled by the effects of surface forces acting across the crack ahead of the tip. The critical stress intensity factor is the value that the surface forces are no longer able to cancel.

When cancellation occurs and the stresses at the contact edge become finite, the edge geometry no longer has the classical parabolic shape $h \sim x^{1/2}$ but becomes a cusp $h \sim x^{3/2}$. Classic papers by Schapery (1975, 1989) explain how this leads to finite strain rates and so to a dependence on loading rates. More detailed analyses of the behavior of opening and closing cracks, using particular surface force laws, have been published by Hui and his collaborators (Hui et al. 1998; Lin et al. 1999), Greenwood and Johnson (1981), Greenwood (2004), and Barthel (1999), confirming Schapery's concept of an apparent surface energy, proportional to the true surface energy, but with a speed-dependent factor that can raise the surface energy by a factor of up to $1/k$ for an opening crack, or reduce it by a factor approaching k for a closing crack. (Recall that k may be as small as 0.01). Figure 4 shows the relation between the apparent surface energy and the crack (contact edge) speed for a three-element solid when the surface forces acting across the gap are represented by the Dugdale model [$\sigma(h) = \sigma_0$ when $h < h_0$; $\sigma(h) = 0$ when $h > h_0$; $\Delta\gamma = \sigma_0 h_0$].

How is this to be connected with the viscoelastic behavior of the sphere itself? A relatively easy case is of some experimental importance: that of a notionally elastic solid such as poly (dimethyl) siloxane (PDMS), which in fact behaves viscoelastically at very high rates of strain. For slow loading, these high strain rates occur only at the contact edge. The bulk of the sphere behaves elastically. Accepting the idea that the apparent surface energy gives a true representation of the behavior, then the JKR equation becomes $P = \frac{4}{3}\frac{E'_\infty}{R}a^3 - \sqrt{8\pi E'_\infty \beta\,\Delta\gamma\, a^3}$, where $\beta \equiv \mathcal{G}/\Delta\gamma$ is the ratio of the apparent surface energy to the true surface energy. This may be rewritten

$$\beta = \left[4E'_\infty a^3 - 3RP\right] \Big/ 6R\sqrt{2\pi E'_\infty \Delta\gamma\, a^3} \qquad (7)$$

and since $\beta = f(v)$ where $v = V\,da/dt$ with V a material constant, this is a differential equation that can be used to follow the increase and decrease of the contact radius by inserting the desired variation of the load P. The approach can, in this case, be found simply from $\delta = a^2/R$. Figure 5 shows how this is used to simulate experimental results by Wahl et al. (2006) in which the load was (slowly) linearly



Adhesive Contact of Inelastic Bodies, Fig. 5 Viscoelastic loading and unloading, with superimposed oscillatory load

increased and decreased, but during unloading a relatively high frequency oscillatory load applied.

Two points may be noted concerning the loading/unloading curves: there is hysteresis as the contact radius on loading is less than the elastic JKR value (calculated using the relaxed modulus), while on unloading it is always larger. The pull-off load is larger than the JKR value, which in these units is $P = -0.5$.

More surprisingly, when an oscillatory load is added, the contact radius largely ignores the oscillation. The crack-tip strain rate is high enough to "pin" the contact edge.

The general case, when both the sphere and the crack tip behave viscoelastically. is more difficult. (If the creep function has a *single* time constant, the very different strain rates in the bulk material and at the crack tip exclude this: either the sphere or the crack will be elastic. But real materials have a whole spectrum of time constants.) Hui et al. (1998) show that during *loading* equation (7) above should be replaced by

$$\beta = \left[4E'_\infty a^3 - 3R\{\Phi \bullet P\}\right] \Big/ 6R\sqrt{2\pi E'_\infty \Delta\gamma\, a^3},$$

$$\text{where } \Phi \bullet P \equiv \int_{0_-}^{t} \Phi(t-\tau)\frac{dP}{d\tau}d\tau$$

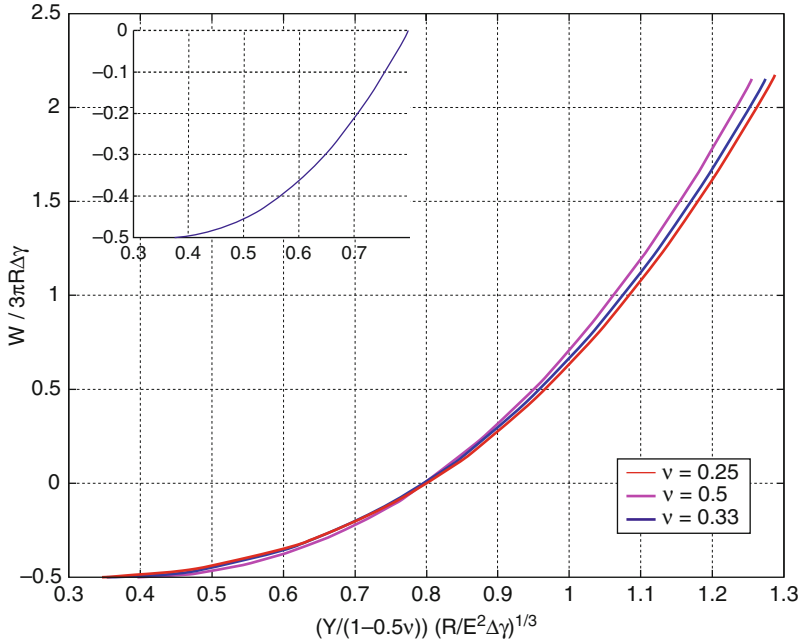
corresponding to (4) above, and this is generally accepted. The unloading process is still unresolved. A convincing analysis by Hui et al. (1999) is challenged, and an alternative offered, by Barthel (1999). Which is correct, and whether either adequately describes real behavior remains to be seen.

It must always be kept in mind that the simple idea that the *real* surface energy needed to make contact will be recovered on breaking contact is far from universally true. As noted earlier, surfaces react while in contact; bonds form and molecules entangle. Viscoelasticity will make the apparent surface energies of bonding and fracture very different, but the real surface energies may themselves be very different.

Plastic Adhesive Contacts

Introduction

Irreversibility in the adhesive contact of spherical surfaces can arise through plastic deformation of the spheres as well as viscoelasticity. This section will be concerned with plastic materials whose deformation is independent of rate. Most practical problems involving adhesion arise at small-scale contacts such as minute particles, or in the contact of rough surfaces, where individual asperities generally deform plastically and adhere, as in the Bowden and Tabor model. The development of micro/nanoscale instruments such as the atomic force microscope has stimulated the need for an analysis of plastic adhesive contacts.



Adhesive Contact of Inelastic Bodies, Fig. 6 Load for (subsurface) first yield. For subsurface yield, the effect of Poisson's ratio is not fully accounted for by using the plane strain modulus $E' \equiv E/(1-\nu^2)$

Adhesive Contact of a Sphere and a Plastically Deforming Half-Space: Loading

When an elastic-perfectly plastic plane surface is indented by a rigid sphere, contact is at first elastic, and, if adhesion is present, well-described by the JKR theory. This enables the internal stress field to be calculated, and, as in a non-adhesive Hertzian contact, the material first yields on the axis well below the surface. Figure 6 shows how the load for first yield depends on the material properties. (The strong dependence on Poisson's ratio should cause no surprise. Although often overlooked, a similar dependence is there in the Hertz theory.) It will be seen that plastic yield can occur *even at zero load*. It seems clear that the fully plastic state, in which the contact pressure reaches the hardness $H \approx 3Y$, can also be attained at zero load. Maugis and Pollock (1984) suggest the general relation that full plasticity is reached when $P + 2\pi\,\Delta\gamma\,R = \pi\,a^2 H$, provided that the radius a exceeds the value needed for full plasticity in the non-adhesive case, $a_p \approx 60RY/E$. This implies that fully plastic indentation can occur under zero load if $\Delta\gamma \geq 6000\,RY^3/E'^2$. (Johnson (1985) suggests a lower value $a_p \approx 30RY/E'$, which leads to $\Delta\gamma \geq 1500\,RY^3/E'^2$, but how exactly is attaining full plasticity to be identified?) The factor R is perhaps the most significant feature; this will happen for *small-scale* contacts.

Unloading from the Fully Plastic State

Unloading is usually an elastic process leaving the surface with a residual deformation, which is often assumed to be spherical and so of radius R_e (More accurately, this is the reciprocal of the relative curvature of the indenter and the recovered profile) found from the Hertz equation $a^3 = \frac{3PR_e}{4E'}$ for the corresponding load and contact radius:

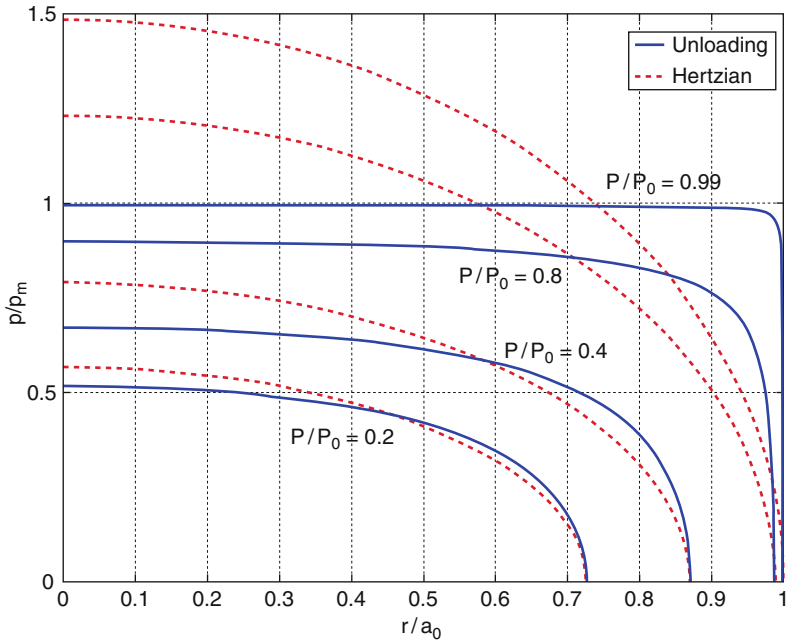
$$R_e = \frac{4E'\,{a_0}^3}{3P} = \frac{4E'\,a_0}{3\pi\,H} \tag{8}$$

Thus, when the load is reduced from an initial value $P_o(=\pi\,a_0^2 H)$, where a_o is the contact radius at the end of plastic loading, two possibilities exist:

1. The contact peels apart from the edge as in a JKR elastic contact, referred to as "brittle" separation, or
2. The load $T_p = \pi\,a_0^2 H = -P_0$ is less than the JKR pull-off load $T_c = (3/2)\pi R_e\,\Delta\gamma$, the surfaces adhere, and failure is by plastic flow (ductile separation). Approximating the recovered profile to a segment of a sphere as above gives the condition for ductile separation as (Johnson (1985))

$$H^2 \leq (2/\pi)E'\Delta\gamma/a_o \tag{9}$$

In fact, the profile remaining after fully plastic loading will not be spherical, but that found when the



Adhesive Contact of Inelastic Bodies, Fig. 7 Pressures during unloading from fully plastic contact

base recovers from a uniform pressure over the contact circle. It can be shown (see Johnson (1985) or Timoshenlo and Goodier (1951)) that the radius of curvature then decreases steadily from the value $R'' = E' a_o / H$ at the center of the recovered profile $(r/a = 0)$ to zero at the periphery (infinite curvature), so the condition for ductile separation depends on where the separation takes place.

This has been analyzed in detail by Mesarovic and Johnson (2000) using "rigid punch decomposition" in which the initial (uniform) pressure distribution is represented as an assembly of rigid flat-ended punches (Hill and Storakers (1990)) according to

$$\begin{aligned} f(\xi) &= -\frac{2}{E'}\frac{\partial}{\partial \xi}\int_{\xi}^{a_0} \frac{r\,p(r)}{\sqrt{r^2-\xi^2}}\,dr \\ &= -\frac{2p_m}{E'}\frac{\partial}{\partial \xi}\int_{\xi}^{a_0} \frac{r}{\sqrt{r^2-\xi^2}}\,dr \\ &= \frac{2p_m}{E'}\frac{\xi}{\sqrt{a_0^2-\xi^2}} \qquad 0 \le \xi \le a_0 \end{aligned}$$

Then, the punches with radii $0 < r < a$ give a pressure distribution

$$p(r) = \frac{E'}{\pi}\int_r^a \frac{f(\xi)\,d\xi}{\sqrt{\xi^2-r^2}} = \frac{2p_m}{\pi}\,asin\sqrt{\frac{a^2-r^2}{a_0^2-r^2}} \qquad (10)$$

and a load

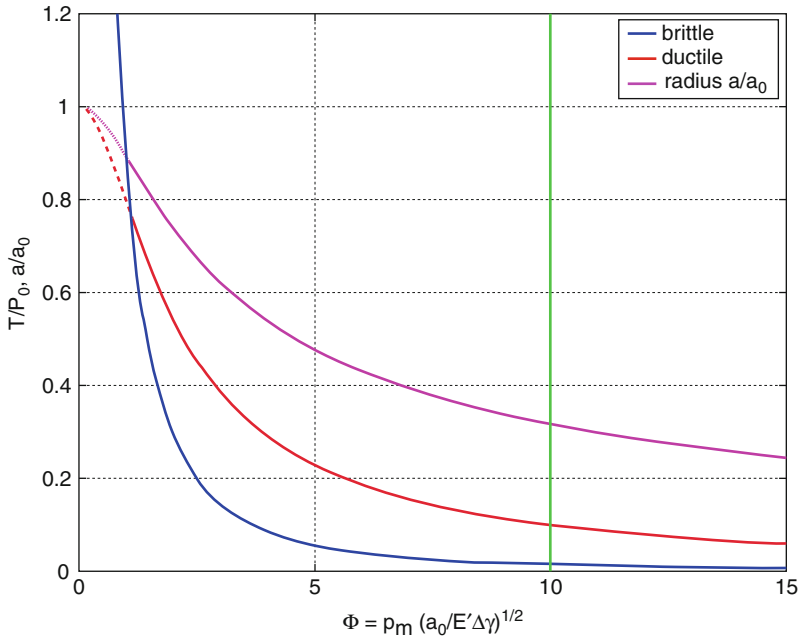
$$P = 2p_m a_0^2\left[asin(\varsigma) - \varsigma\sqrt{1-\varsigma^2}\right] \quad \text{where } \varsigma \equiv a/a_0 \qquad (11)$$

This non-adhesive unloading is shown in Fig. 7. The unloading pressures are very different from the Hertzian pressures frequently assumed.

Adhesion may now be introduced by the JKR procedure. A Boussinesq "lift" is applied over the current contact radius $r = a$ to give a stress intensity factor there as required by fracture mechanics theory, i.e., tensile pressures $p(r) = -p_1\Big/\sqrt{1-r^2/a^2}$ where $p_1 = \sqrt{2E'\Delta\gamma/\pi\,a}$ are added, contributing a load $-\sqrt{8\pi E'\Delta\gamma\,a^3}$. Thus, the total force (divided by the original load $P_0 = \pi\,a_0^2 p_m$) becomes

$$\begin{aligned} \frac{T}{P_0} &= \frac{2}{\pi}\left[\sqrt{2\pi\,\Phi}\,\varsigma^{3/2} - \left\{asin(\varsigma) - \varsigma\sqrt{1-\varsigma^2}\right\}\right] \\ &\text{where } \Phi \equiv p_m\left(\frac{a_0}{E'\,\Delta\gamma}\right)^{1/2} \end{aligned} \qquad (12)$$

The pull-off force is found by minimizing this, which is when $\Phi^2 = \frac{9\pi}{8}\frac{1-\zeta^2}{\zeta^3}$. Figure 8 shows the resulting pull-off forces. But at a given contact radius a there is also the possibility of a ductile failure with a force $T = \pi\,a^2 H$. And, of course, failure will occur at the lesser of the two forces.



Adhesive Contact of Inelastic Bodies, Fig. 8 Pull-off forces after loading to fully plastic state. Care in interpretation of the ductile failure region is required, see text. More generally, all the diagram to the *left* of the *green line* $\Phi = 10$ is uncertain because of possible effects of local yielding at the contact edge

Note that the ductile pull-off forces are related to the contact radius, but the contact radius can only be found from the parameter Φ in the brittle pull-off range. Thus, for $\Phi = 1$, all that can be said is that a brittle failure would occur at $a/a_0 \approx 0.9$, $T/P_0 \approx 0.9$. But a ductile failure at this load would already have occurred at $a/a_0 \approx \sqrt{0.9} = 0.95$.

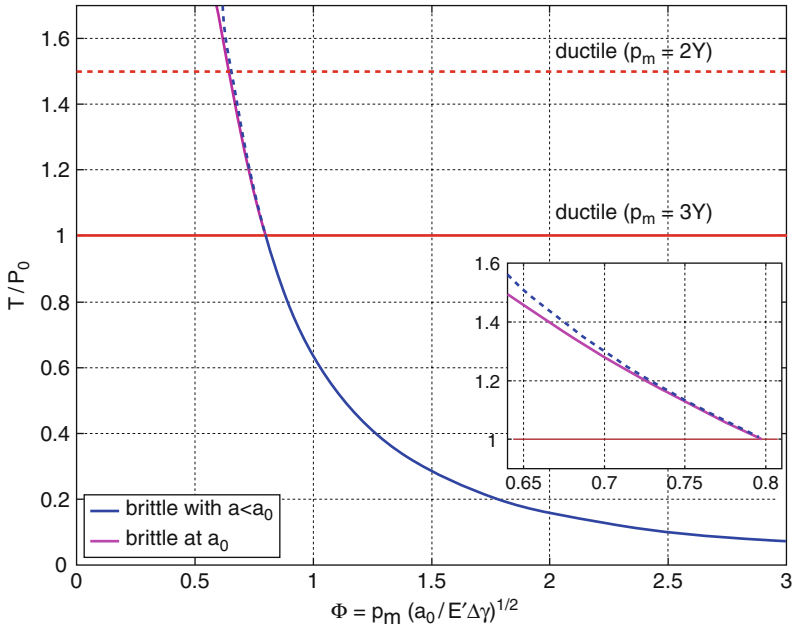
Small-Scale Yielding

It may be objected that, in an adhesive contact, it is futile to study failure on the axis when there are already infinite stresses at the contact edge. Do these not lead to plastic flow? This is a regular problem in elastic fracture mechanics, and the argument there is that although the theoretical stresses $N/\sqrt{x}$ near the crack tip clearly cannot occur, and must be relieved by plastic flow or some equivalent mechanism, the stresses outside the local plastic zone still do follow this law, so that the local yielding may be ignored. The size of the local zone may be estimated by $N/\sqrt{x} = Y$, or $x = E'\Delta\gamma/\pi\, Y^2$. This will be small compared to the contact size (and so may be ignored) when $x << a$.

This becomes $\Phi^2 >> \frac{9}{\pi\varsigma}$, or using the $\Phi(\varsigma)$ relation, $\varsigma^2/(1-\varsigma^2) << \pi^2/8$, which may be interpreted as $\varsigma < 0.3$, or $\Phi > 10$. Thus, much of Fig. 8 is uncertain. Only (oddly) for rather small contacts is the local yielding at the contact edge realistically "small-scale."

The Cohesive Zone Solution

The singular stresses of JKR analysis may be removed by a "cohesive zone" method, often referred to as a Dugdale-Barenblatt method, and with similar results, although the motivations are somewhat different. The Dugdale approach is to consider the head of the elastic region as the crack tip, and assume that ahead of the crack tip there is a tensile yield stress Y trying to close the crack. The Barenblatt concept is that molecular forces act ahead of the crack tip across a real gap and depend on the law relating molecular forces to the gap across which they act. But since Barenblatt's molecular forces are often replaced by a constant limiting stress σ_0 for mathematical convenience, the distinction is blurred. Mesarovic and Johnson use the Dugdale approach but take the limiting stress to be an independent material constant σ_0, thus, introducing a second governing parameter $S \equiv \sigma_0/Y$ into the analysis. It appears, however, that provided $S > 1$, the changes are minor, and perhaps only for ceramics is $S < 1$ and complications arise. Even then the pull-off forces are almost unchanged.



Adhesive Contact of Inelastic Bodies, Fig. 9 Coefficient of adhesion T/P_0. After elasto-plastic loading, there is a narrow range of Φ over which brittle failure at the initial contact radius can occur

Adhesion After Elasto-Plastic Loading

Maugis and Pollock (1984) use the simple Hertzian unloading model described earlier to obtain the effective radius of curvature of (8), and on this they base a full analysis of unloading from both the fully plastic and the elasto-plastic states. They use an empirical equation, based on the spherical cavity model of elastic–plastic deformation (see Johnson (1985)), but modified to give a better fit with experiments:

$$\frac{p_m}{Y} = 1.1 + 0.58 \ \ln\left(\frac{E^* a}{2.3YR}\right) \qquad (13)$$

to describe the growth in mean pressure p_m from elastic, through elastic–plastic to fully plastic. When $\overline{p} = 3Y$, full plasticity is reached and no further increase in p_m takes place. Maugis and Pollock introduce a parameter ϕ defined by $\phi \equiv p_m \left(\frac{\pi a_0}{2E'\Delta\gamma}\right)^{1/2}$ differing from the parameter Φ introduced earlier by the factor $\sqrt{\pi/2}$ but, more importantly, differing in that p_m is now the elasto-plastic mean pressure while before it was always the fully plastic mean pressure $H = 3Y$.

Figure 9 shows their results for the coefficient of adhesion T/P_0. There is a minor difference for unloading from the fully plastic state, in that Maugis and Pollock find only the two possibilities: immediate ductile yielding at the initial contact radius, or brittle separation at a reduced radius. (In contrast, Mesarovic and Johnson suggest that ductile failure at a (slightly) reduced radius can occur.) However, for elasto-plastic loading, immediate ductile failure would give $T/P_0 = \pi a_0^2 H/\pi a_0^2 p_m > 1$, and this provides room for the third mode (only really visible in the insert), brittle failure at the initial radius a_0.

Note that Maugis and Pollock give an explicit condition for the change from ductile to brittle failure. In the present notation,

$$\text{for ductile failure, } \Phi > \sqrt{8/\pi} \,/\, (1 + H/p_m) \qquad (14)$$

It will be seen that if $p_m = H$, this reduces to (9).

Discussion

Maugis and Pollock note that it is natural to base unloading analyses on the initial contact radius, but, at low loads, quite wrong to assume this to be given by $a_0^2 = P_0/\pi p_m$. Adhesion operates during loading as well as during unloading, and may lead to a fully plastic contact even at zero load. A prediction of the pull-off force based on the initial contact area may still be useful, but the coefficient of adhesion would then be infinite!

It follows that throughout this work, P_0 should not be taken to be the initial load, but as the quantity calculated as $\pi a_0^2 p_m$ from estimates of the contact radius a_0 and the contact pressure p_m.

The unloading process is governed by the parameter $\Phi \equiv p_m \left(\frac{a_0}{E' \Delta\gamma}\right)^{1/2}$, and only when this is small ($\Phi < 2?$) will adhesive effects be important. Clearly a large value for the surface energy will help, but the really important region is when the initial contact radius a_0 is small.

Finally, it must be emphasized that whatever the true surface energy, the effects of contamination or surface roughness may reduce the effective value to almost zero.

Cross-References

► Basic Concepts in Adhesion Science

References

E. Barthel, Modelling the adhesion of spheres when the form of the interaction is complex. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 149 (1999), pp. 99–105

J.A. Greenwood, The theory of viscoelastic crack propagation and healing. J. Phys. D: Appl. Phys. **37**, 2557–2569 (2004)

J.A. Greenwood, K.L. Johnson, The mechanics of adhesion of viscoelastic solids. Philos. Mag. **A43**, 697–711 (1981)

R. Hill, B. Storakers, A concise treatment of axisymmetric indentation in elasticity, in *Elasticity: Mathematical Methods and Applications*, ed. by G. Eason, R.W. Ogden (Ellis-Horwood, Chichester, 1990), pp. 199–210

C.-Y. Hui, J.M. Baney, E.J. Kramer, Contact mechanics and adhesion of viscoelastic spheres. Langmuir **14**, 6570–6578 (1998)

K.L. Johnson, *Contact Mechanics* (Cambridge University Press, Cambridge, 1985)

Y.Y. Lin, C.Y. Hui, J.M. Baney, Viscoelastic contract, work of adhesion and the JKR technique. J. Phys. D: Appl. Phys. **32**, 2250–2260 (1999)

D. Maugis, H.M. Pollock, Surface forces, deformation and adherence at metal microcontacts. Acta Metall. **32**(9), 1323–1334 (1984)

S.Dj. Mesarovic, K.L. Johnson, Adhesive contact of elastic–plastic spheres. J. Mech. Phys. Solids **48**, 2009–2033 (2000)

R.A. Schapery, A theory of crack initiation and growth in viscoelastic media. Int. J. Fract. **11**, (Part I) 141–159; (Part II) 369–388; (Part III) 549–562 (1975)

R.A. Schapery, On the mechanics of crack closing and bonding in linear viscoelastic media. Int. J. Fract. **39**, 163–189 (1989)

S.P. Timoshenlo, J.N. Goodier, *Theory of Elasticity* (McGraw-Hill, New York, 1951)

T.C.T. Ting, The contact stresses between a rigid indenter and a viscoelastic half-space. ASME J. Appl. Mech. **35**, 845–854 (1966)

K.J. Wahl, S.A.S. Asif, J.A. Greenwood, K.L. Johnson, Oscillating adhesive contacts between micron-scale tips and compliant polymers. J. Coll. Interface Sci. **296**, 178–188 (2006)

Adhesive Contacts

► Contacts Considering Adhesion

Adipose Tissue

► Fat Pad Lubrication Properties

ADORE, Advanced Dynamics of Rolling Bearings

► Analytical Modeling of Rolling Bearings

Adsorption

► Solid–Liquid Bi-phase Lubricating Coatings

Aerostatic Bearings

► Hydrostatic/Hybrid Gas Bearings

Air Bearing

► Air Bearing Diagnosis

Air Bearing Diagnosis

Jianfeng Xu[1], Yiao-Tee Hsia[1], Gang Sheng[2]
[1]HDD H/M Tribology & Mechanical Integration, Western Digital Corporation, San Jose, CA, USA
[2]Department of Mechanical Engineering, University of Alaska, Fairbanks, AK, USA

Synonyms

Air bearing; Acoustic emission; Fly height interferometry

Definition

Diagnostics of nanometer spacing air bearing uses the technique of monitoring multiple interface parameters

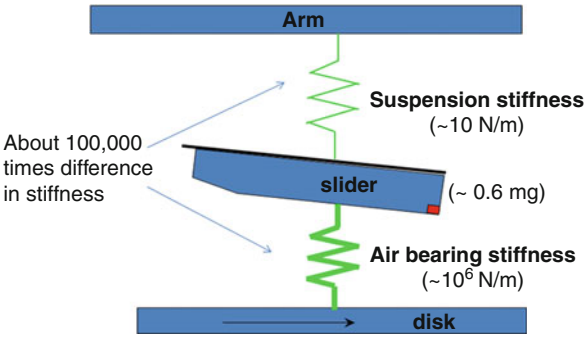
to explore the feasibility and durability of slider air bearings used for contemporary hard disk drives.

Scientific Fundamentals

In hard disk drives, a slider at the end of a suspension flies over a fast-rotating disk due to the squeezed air bearing film. The air bearing film is very thin; the thinnest clearance is only a few nanometers in current hard disk drives. The slider houses magnetic read/write heads to conduct read/write operations on the magnetic disk. Thus, a magnetic recording is available. In order to increase the magnetic recording density, the flying height should be low and stable. When the flying height is too low, slider-disk contact may happen, and damage to the head and disk is expected. When the disk is rotating, the amplitudes of disk vibrations are normally in the micrometer level. In addition, the installation tolerances of the suspension assembly and the disks are also in the micrometer level. The trick is to allow a slider to follow the disk to obtain the nanometer spacing, i.e., a nanometer flying height, for magnetic recording in the air bearing film between the slider and the disk. Figure 1 illustrates the schematic mechanical model of the slider and the disk. The huge stiffness difference between the suspension and the air bearing makes the micrometer level vibrations and fabrication tolerances shift little in the flying height. The slider is somewhat "glued" to the disk and follows the disk closely, avoiding slider-disk contact. In this way, nanometer-level flying height of an air bearing slider can be obtained for magnetic recording. Figure 2 illustrates the schematic of air bearing with 10 nm spacing, aiming at pushing the recording density towards 1 Tb/in^2 and beyond. The reduced head-disk spacing leads to more severe contact and increases the chances of damage to the recording transducer during both start/stop or load/unload, as well as flying cases (Menon 2000).

The Factors Influencing Feasibility and Durability of an Air Bearing with Nanometer Spacing: A list of the basic factors influencing feasibility and durability of nanometer spacing air bearing interface is as follows:

1. Fly height (FH) and its modulation, take-off height, glide height, instability near the glide avalanche point affected by lubricant.
2. Contact, head-disk interaction mediated by van der Waals force and meniscus formation, lubricant pickup, lubricant thickness modulations.
3. Contamination, particle buildup.
4. Tribocharge.
5. Wear, collision, and disk damage, lubricant degradation.



Air Bearing Diagnosis, Fig. 1 Mechanical model of the slider and the disk

Diagnostic Approaches

In the following, the basic diagnosis approaches are described, which include the capacitance method, optical method, acoustic emission method, method based on read/write signal, particle measurement, embedded sensor method, and tribocharge characterization.

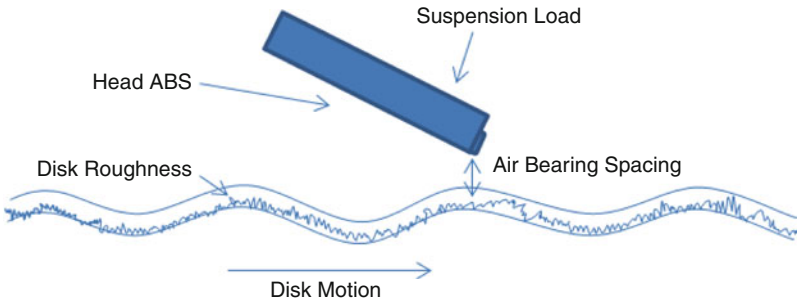
Capacitance Method

The capacitance method was used to conventionally measure FH of an air bearing slider (Briggs and Herkart 1971). As a slider flies over a disk surface, there is a capacitance between the slider and the disk. The capacitance value is related to the flying height. By measuring the capacitance, the data related to the flying height can be obtained. The method is relative easy to apply for measurement, but it normally cannot provide the flying height of a specified point on slider, as the capacitance is related to the flying height of the whole air bearing surface of the slider. In order to measure the flying height of the specified point, the slider needs to be specially fabricated. For example, by fabricating three small capacitances to the air bearing surface, the flying heights of the three specified points could be measured independently. This method is very difficult for characterizing current air bearing sliders.

Optical Methods

Many optical measurement methods are widely used to diagnose air bearing interface. The popular ones include the light interference method, light intensity/phase method, laser Doppler vibrometer (LDV) method, and the surface reflectance analyzer (SRA) method.

The white light interferometry method has been the industry standard to measure FH of air bearing sliders, but it has poor repeatability for sub-10 nm FH measurements. The light intensity/phase method is more sensitive for characterizing flying height. The laser phase detection



Air Bearing Diagnosis, Fig. 2 Schematic of the interface with air bearing slider and disk with nanometer spacing

FH tester has been applied. It performs the on-flying calibration and can improve the repeatability of the FH test significantly. The weakness of the above two optical methods is that a glass substrate disk without lubricant has to be used in the test, which is somewhat different from the real media disk (Suk et al. 1991).

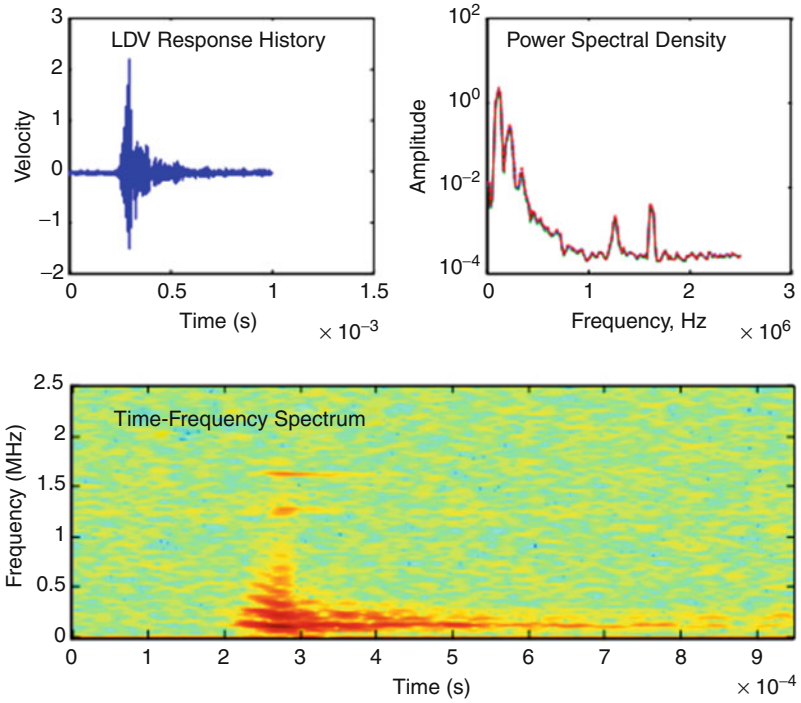
The LDV method is normally used to detect the vibration speed of an air bearing slider by casting the laser beam to the back of slider. The advanced scheme can be used to detect not only the flying height variation, but also the pitch and roll angles with multi beams and multi channels. The response of an air bearing slider passing a specified bump is usually applied for contact validation (Briggs et al. 1990; Xu et al. 2007). The LDV method may not be sensitive to light contact between the slider and the disk.

Example: The slider used was a 30% pico (1 mm 0.8 mm 0.3 mm) slider. The disk has 1.5 nm carbon overcoat and 1.5 nm lubricant, roughness Ra = 0.4 nm. The slider was designed to have 8 nm FH at the trailing edge, which was met at a test radius of 40 mm on the spinning disk of 3,600 RPM. The output velocity signal of the LDV was fed into a high-pass filter with a cutoff frequency of 20 kHz. The resonant modes of the slider air bearing are in the range of 30–200 kHz from its specification. The major natural frequencies of the slider body are in the range of 1–2 MHz using FEM analysis. The velocity output was fed into a band-pass filter with the frequency range from 2 to 300 kHz for the air bearing analysis. Alternatively, the band-pass filter was removed to analyze the high-frequency signal for slider body resonance due to slider-disk contact. To validate the contact test, an experiment was conducted by allowing the 8 nm FH slider to fly across a specified bump on disk with 10 nm height and 10 μm radius. The semi-sphere bump is fabricated using a laser process and cubic bump fabricated using etching techniques. The slider is likely to make contact with the bump at properly low disk spinning speed in the experiment. Figure 3 shows the LDV measured response time history, spectrum peak hold, and time-frequency spectrum of response of an air bearing slider to bump. The figures show that both air bearing resonant modes (below 500 kHz) and slider body resonant modes (between 1 and 2 MHz) were excited during the bump excitation. It indicates that the slider made contact with the bump.

LDV also allows one to obtain the "in-plane" motion for a slider from the torsion mode and sway mode of the suspension assembly, and the "out-of-plane" motion for slider from torsion and bending modes.

SRA is in principle a high-speed scanning ellipsometry. It is optical equipment for studying and quantifying the composition and thickness of thin films on surfaces, such as lubricant layers or carbon coatings on hard disks. It can be used to locate and quantify tribological parameters, such as carbon wear and lubricant buildup, at the air bearing disk interface. The disk damage caused by a variety of head/disk interactions can be characterized by quantitatively comparing the damage to the disk surface during continuous running operations like a start-stop test or load/unload. The SRA system can be used to distinguish and quantify these various types of damage like local carbon wear, lubricant effects, and other abnormal effects like the presence of debris (Vurens et al. 2000).

SRA gives film thickness information on the carbon and lubricant film over the disk surface by measuring the polarization state of light reflected from the disk surface. This instrument is particularly useful in measuring wear of the disk carbon film as well as for studying the redistribution of lubricant by the head disk interaction. The instrument can also be used to look at defects, because, in addition to the specularly reflected light, light scattering is also measured. This scattered light detection gives the instrument a way to quantify changes on the disk surface.



Air Bearing Diagnosis, Fig. 3 LDV-measured response time history, spectrum peak hold, and time-frequency spectrum of response of an air bearing slider to bump

The SRA is sensitive to optical properties n and k and the thickness of thin films. Here, n is the index of refraction and k is the extinction coefficient. A low-intensity visible laser beam is reflected from the surface of a disk at a fixed angle and the polarization properties (P-polar reflectivity and phase contrast) of the reflected beam are measured at high data rate (up to 30 MHz) and high accuracy. The polarization of the reflected beam is a function of the refractive index, the absorption coefficient, and the thickness of the thin film from which the beam is reflected. The SRA maps the optical properties of the reflected beam on the disk surface. The maps are then analyzed to give a measure of the thickness uniformity and phase (composition) uniformity. Although the SRA does not measure the absolute thickness value of a thin film, it can give the thickness uniformity values with accuracy on the order of 0.05 nm.

Acoustic Emission Method

The acoustic emission (AE) method is often used to detect slider-disk contact. When a slider is in contact with the disk, acoustic emission is detected and amplified by a sensor mounted on the arm. It is a highly efficient approach to detecting and investigating slider dynamics and slider-disk interaction. An acoustic emission sensor is generally a piezoelectric element, which transforms the particle motion produced by an elastic wave into an electrical signal. Some AE sensors are stand-alone assemblies that can be mounted or glued on an arm or suspension; other AE sensors are piezoelectric chips (PZT sensor) that are usually integrated onto slider body. A PZT sensor is much closer to the contact point, so it is more sensitive to the slider-disk contact (Sheng and Liu 1999; Khurshudov and Talke 1998).

The acoustic signal (AE) method is the typical method used for the detection of slider–disk contact. In addition to detecting disk-bump and slider-asperity contact, AE signal variation can even reflect the slight slider-disk interaction in slider thermal protrusion for thermal flying height control slider, due to the electrical current applied to the writer that drives the head to contact the disk surface (touch-down) or leave disk surface (take-off). Hysteresis can be observed in take-off and touch-down testing. For other specific applications, AE signal change can be an indicator of particle contamination in an interface. When particles intrude

into an interface, it generate AE peaks and the duration of the signal for the agglomerated soft particles tend to be longer compared with the signal observed for hard particles. Using these results, AE signal analysis can be effectively used as a monitoring method for contamination at air bearing interface. The AE approach has been widely used to quantify take-off height, glide height, and instability near the glide avalanche point affected by lubricant.

In Situ Method Based on Read/Write Signal

The read/write signal-based method has been used to detect flying height as an in-situ approach. As the read signals are inversely proportional to the flying height, the flying height can be calculated by detecting the read signals. The method is easy, practicable, and accurate, and is conducted on the real media disk. However, the method can only detect the relative flying height. If it is calibrated by other methods, such as the optical method, the relative flying height can then be converted to the detailed flying height values. The readback signal-based in-situ FH testing, triple-harmonic method was used to measure the real-time FH signal and to observe the flying stability of slider during the head heater excitation (Yuan et al. 2002).

Particle Measurement

To characterize particle effect, the number of background and generated particles residing in the chamber of the air bearing interface can be minimized/maximized by blowing dry and clean or polluted compressed air into the chamber after passing the air through a water trap, diffusion drier, and HEPA (high efficiency particulate air) filter. The environmental particle changes due to contamination can be monitored by using a condensational particle counter with varied level 1–1,000 particles/m3 (Chung et al. 2004).

Embedded Sensor Method

The embedded sensor method has been used in air bearing diagnosis. This method utilizes thermal energy and/or force associated with intermittent head-disk contacts. One approach includes a frictional force sensor embedded to the backside of a slider and attached to a suspension. In the event of head-disk contact, the lateral motion of the slider will be excited and sensed by the force sensor. The signal is detected by specifically designed electronics and can be used to determine the contact force and contact duration.

Another common approach is to use a heat sensor embedded into the slider read/write element. When contact occurs in the head-disk interface, temperature fluctuation around the contact surfaces often shows up. By monitoring the temperature change of the heat sensor, the event of contact could be determined.

Tribocharge Characterization

One important effect associated with intermittent head-disk contacts/friction is the tribocharge/tribocurrent phenomenon, which is the process whereby a charge exists on a material after the parting of solid/solid or solid/fluid contacts. The magnitude of the final charge will actually be the result of two processes: the charge transfer that occurs during the sliding, and the charge backflow that occurs as the materials are separated. The tribocharge buildup on a magnetic recording slider surface is important because the buildup can increase slider-disk interaction level, induce lubricant degradation, or accelerate contamination on the slider-disk interface, as well as cause electrostatic discharge damage. Tribocharge level in a slider-disk interface depends mainly on the electric capacity of the material involved and on the relative speed of the slider to the disk. The current and voltage generated between the slider and the disk can be measured using an electrometer. The electrometer can be connected to the suspension assembly. For detecting the tribocharge/tribocurrent signals in the slider, the suspension needs to be electrically isolated from other parts. The tribocharge/tribocurrent signals are measured at the same time. The tribocharge decays while there is no interaction between the slider and the disk. The decay of tribocharge is inversely proportional to the square root of time (Kurita et al. 2005; Park et al. 2004; Lee et al. 2007).

Key Applications

The self-regulating air bearing of the head-disk interface in the HDD is the most important application of gas lubrication, which provides the separation between head and disk and offers reliability to hard disk drive systems. Air bearing diagnosis technology has been the key technology for the development of air bearing sliders and interfaces over the last 50 years. The head-disk spacing has been reduced from more than 5,000 nm to less than 5 nm. Currently, thermal fly-height control (TFC, or dynamic flying height (DFH)) technology has been widely applied in air bearing sliders. TFC employs a heating element for active flying height control. When power is supplied to this element, the slider thermally expands locally at the trailing portion that protrudes the read-write transducers to be closer to the disk. Because of the smallness of the close-approach region, the

destabilizing forces are minimized. TFC has been developed for data densities greater than 1 Tbit/in^2. The further development of magnetic recording for 10 Tbit/in^2 requires magnetic spacing is no more than 2.5 nm. Within this magnetic spacing there must be a slider-disk spacing less than 0.5 nm, which needs a intermittent/continuous contact interface. In the development of this kind of interface, to overcome the difficulty of instability, reliability, wear, and lubricant, air bearing diagnosis technology has been widely used as an enabling technology for testing, characterization, and design validation (Sheng 2011).

Cross-References

- ► ABS Designs
- ► Disk Roughness and Defect Monitoring
- ► Hard Disk Drive (Box Level)
- ► Suspension Assembly for Hard Disk Drive

References

G.R. Briggs, P.G. Herkart, Unshielded capacitor probe technique for determining disk memory ceramic slider flying characteristics. IEEE Trans. Magn. **MAG-7**(3), 418–421 (1971)

C.A. Briggs et al., The dynamics of "micro" sliders using laser doppler vibrometry. IEEE Trans. Mag. **MAG-26**(5), 2442–2444 (1990)

K.H. Chung et al., Particle monitoring method using acoustic emission signal for analysis of slider/disk/particle interaction. Tribol. Int. **37**, 849–857 (2004)

A. Khurshudov, F.E. Talke, A study of sub-ambient pressure tri-pad sliders using acoustic emission. ASME Trans. J. Tribol. **120**, 54–59 (1998)

M. Kurita et al., Flying-height reduction of magnetic-head slider due to thermal protrusion. IEEE Trans. Mag. **41**(10), 3007–3009 (2005)

D.Y. Lee et al., Effect of relative humidity and disk acceleration on tribocharge build-up at a slider–disk interface. Tribol. Int. **40**, 1253–1257 (2007)

A.K. Menon, Interface tribology for 100 Gb/in2. Tribol. Int. **33**, 299–308 (2000)

H.S. Park, J. Hwang, S.H. Choa, Tribocharge build-up and decay at a slider-disk interface. Microsyst. Technol. **10**, 109–114 (2004)

G. Sheng, Sensing and identification of nonlinear dynamics of slider with clearance in sub-5 nanometer regime (Invited). Adv. Tribol. (2011). doi:10.1155/2011/282839. Hindawi, ID 282839

G. Sheng, B. Liu, A theoretical model of slider-disk interaction and acoustic emission sensing process for studying interface phenomena and estimating unknown parameters. Tribol. Lett. **6**, 233 (1999)

M. Suk et al., Comparison of flying height measurement between multi-channel laser interferometer and capacitance probe slider. IEEE Trans. Mag. **MAG-27**(6), 5148–5150 (1991)

G.H. Vurens et al., Tribology applications of surface reflectance analyzers: optical characterization of the head disk interface. Tribol. Int. **33**, 647–653 (2000)

J. Xu et al., Dynamics of ultra low flying sliders during contact with a lubricated disk. Microsyst. Technol. **13**, 1371–1375 (2007)

Z. Yuan et al., Engineering study of triple-harmonic method for in situ characterization of head-disk spacing. JMMM **239**(1–3), 267–370 (2002)

Aircraft Engine Lubricants

Jun Dong, Cyril A. Migdal, Dale Carr
Chemtura Corporation, Middlebury, CT, USA

Synonyms

Aircraft piston engine oils and turbine engine oils

Definition

Aviation engine lubricants are designed to lubricate, under a wide range of operating temperatures, engine moving parts such as bearings, gears, camshaft, rocker arms, cylinder walls, piston rings, push rods, and sockets and provide additional functionalities for engine cooling, cleanliness, and corrosion inhibition.

Scientific Fundamentals

Lubricants for Aircraft Piston Engines

The majority of piston engines can be classified as one of the two basic types: radial engines or in-line engines. Radial engines may have one, two, or four rings of cylinders, each consisting of three to nine units radially mounted around an axis. There are always an odd number of cylinders in each ring. In-line engines may have one or more banks of cylinders placed horizontally or vertically or in various configurations and accordingly resulting in straight engines, horizontally opposed or so-called flat engines, V engines, and H engines. The bulk of general aviation is powered by flat engines manufactured by Textron Lycoming or Teledyne Continental Motors (Poitz and Yungk 2006).

The earliest piston engines were lubricated with either mineral oils or vegetable oils, particularly castor oil. Castor oil is a triglyceride, but unlike other vegetable oils, it also has a very high content of ricinoleic acid, a fatty acid containing one double bond and one hydroxyl functional group. This unique composition gives the oil excellent boundary lubrication characteristics, while maintaining good low temperature fluidity without the oxidative instability of polyunsaturates that are commonly present in vegetable oils. Mineral oils, on the other hand, were inferior in the area of boundary lubrication and early products generally suffered from poor viscosity-temperature characteristics and low oxidation resistance. Consequently, large-scale replacement of vegetable oils by mineral oils did not take place until the 1930s, when the quality, availability, and cost of mineral oils started to improve.

For many years aircraft piston engines operated satisfactorily on straight mineral oils. Highly refined petroleum base stocks were extensively used owing to their good viscosity-temperature characteristics, better inherent oxidation resistance, and low levels of sulfur species, which are known to form corrosive acids upon reacting with moisture and other combustion by-products. The essentially additive-free, straight mineral based engine oils are defined in greater detail in the current Society of Automotive Engineering (SAE) Standard J1966 (formerly Military Specification MIL-L-6082). Oils of this class are sometimes referred to as running-in or break-in oils for their ability to provide the correct level of lubricant breakdown and controlled cylinder wear to help lap and seal the piston rings. Oils meeting this specification may contain a low level of ashless antioxidant and a pour point depressant for improved high- and low-temperature performance. Ash and deposit-forming additives are prohibited because their degradation products can be insoluble in these oils and, if not controlled, can block the oil passageways in the engine. Also, the motion of the aircraft during operation tends to prevent proper settlement of the solids.

With the advance of ashless dispersants, antioxidants, antiwear, and anti-foam technologies in recent years, the use of performance-enhancing additives in aircraft piston engine oils has been increasing and some of the additives are now permitted in high-performance engine applications. The oils containing dispersants and other additives are also known as "dispersant" oils and are currently specified in the SAE Standard J1899 (formerly MIL-L-22851). Oils of this class are generally superior in terms of providing enhanced low-temperature fluidity, high-temperature stability, better engine cleanliness, better corrosion inhibition, and antiwear protection. As such, oils meeting this specification now represent more than 70% of aircraft piston engine oils sold (Poitz and Yungk 2006).

Lubricants for Aircraft Turbine Engines

There are three major classes of aircraft turbine engines that power the aviation industry. By chronological development from the earliest, they are (1) the turbo-jet engine, in which the whole propulsive force is provided by the jet thrust with a small portion being extracted to drive the compressor and some auxiliary components; (2) the turbo-prop or prop-jet engine, which is similar to a turbo-jet engine but has a large outside propeller placed in front of the engine (a high proportion of the power from the combustion gases is extracted to drive the propeller which provides virtually all the propulsive force to the aircraft); and (3) turbo-fan engines. This class of engine utilizes a front, ducted fan driven by the turbine that generates two streams of airflow. The first stream flows through the turbine core, providing compressed air to burn fuel in the combustion chambers. The second stream is ducted to flow outside the turbine core at a relatively lower velocity. This type of design allows a combined benefit from both turbo-jet and turbo-prop engines for better fuel efficiency at high airspeeds.

When compared with piston engines, the lubrication requirements for turbine engines, especially turbo-jet and turbo-fan engines, are less demanding because of the relatively simple engine design and construction that excludes many moving parts in the combustion chambers. However, aircraft turbines operate at much wider temperature range, typically from −54°C to 200°C or higher (Landsdown 1987), and therefore require the engine oils to have adequate low-temperature fluidity and high-temperature stability. Highly refined mineral oils with low viscosities (typically from 2 cSt to 9 cSt at 100°C) were used in the early development of turbine engines with limited success. It was found that even the best mineral oils lacked the high-temperature stability necessary for high-power and high-performance turbine engines. Major issues included high oil volatility and severe oxidation, which led to excessive oil thickening and the formation of engine sludge and deposits.

The performance requirements for the extreme operating temperatures of turbine engines led to the development of synthetic ester-based oils. Synthetic esters are basically reaction products of carboxylic acids and alcohols. With proper selection of the raw materials, the resulting esters can possess lubrication properties far superior to those of mineral oils. In general, advantages of synthetic esters include (1) good low temperature fluidity, (2) excellent solvency for additives, oxidation products, and contaminants, (3) good thermal-oxidative stability, (4) good deposit control capability, (5) good anti-wear and load-carrying capability, and (6) environmental friendliness. The first generation of synthetic oils used for aircraft turbine engines (Type 1) was based on diesters, usually blends of alkyl adipates, azelates, and sebacates, especially dioctyl sebacate (Table 1). This class of esters generally exhibited very good viscosity index (VI) and pour point characteristics, which allowed these engine oils to operate over a wide temperature range. However, one disadvantage associated with most diesters was their low viscosity (typically of about 2–8 cSt at 100°C). In order to function properly in turbine engines, they required complex esters as thickeners to meet the load-carrying requirements, such as for supporting and transmitting high gear loads. In addition, the thermal stability of diesters was of concern because of the presence of an abstractable hydrogen

Aircraft Engine Lubricants, Table 1 Representative structures of synthetic esters used in aircraft turbine oils

$R_1OOC(CH_2)nCOOR_2$	**Diester** R_1, R_2 = linear, branched, or mixed alkyl chain n = 4 (adipates), 7 (azelates), and 8 (sebacates)
$CH_3CH_2C(CH_2OOCR)_3$	**Trimethyolpropane ester** R = linear, branched, or mixed alkyl chain
$(CH_3)_2C(CH_2OOCR)_2$	**Neopentylglycol ester** R = linear, branched, or mixed alkyl chain
$C(CH_2OOCR)_4$	**Pentaerythritol ester** R = linear, branched, or mixed alkyl chain
$(RCOOCH_2)_3CCH_2$-O-$CH_2C(CH_2OOCR)_3$	**Dipentaerythritol ester** R = linear, branched, or mixed alkyl chain

atom on the beta carbon of the alcohol portion of the molecule, which limited the bulk oil operating temperature to about 170°C. As newer engines operated at higher loads and temperatures, diesters became inadequate to meet the performance requirements, and subsequently most of them have been replaced by the more thermally stable neopentyl polyol esters, including trimethyolpropane esters, neopentylglycol esters, pentaerythritol esters, and dipentaerythritol esters (Table 1). These esters offer excellent low-temperature viscosity, high viscosity index, and very high autogenous ignition temperature characteristics. Depending on the application, modern turbine oils are usually formulated with antioxidants, dispersants, corrosion inhibitors, antiwear, and extreme pressure (EP) additives to enhance performance.

Key Applications

Type of Lubrication System

Most aircraft engines utilize either a wet sump or dry sump type of lubrication system. One of the primary differences between the two systems is the wet sump stores oil in the engine crankcase, while the dry sump uses an external oil tank mounted on or near the engine. The wet sump type is very similar to those used in automotive engines and is commonly employed by the low-output in-line engines usually found in the so-called light or private planes. Larger aircraft engines, especially radial engines, usually operate on the dry sump lubricating system in which returning oil is first collected in a small sump located in the engine and then removed by scavenge pumps and transferred to the main oil tank separate from the engine.

Piston Engine Oils

Piston engine oils meeting both SAE J1966 and J1899 standards must possess important properties so that they can function properly under designated service conditions, in particular, at operating temperatures ranging from −35°C to more than 95°C and at an altitude of 6,000 m. Basic functionalities that a piston engine oil is required to provide include:

- Lubricate cylinders, piston rings, valves, gears, and bearings to reduce friction and wear. Provide additional sealing between piston rings and cylinder walls.
- Provide cooling to engine hot parts and areas.
- Provide cushion to moving parts against shock.
- Protect engine internal parts against corrosion.
- Keep engine interior and moving parts clean and free of dirt, sludge, varnish, and other harmful contaminants.
- Function as hydraulic oil for operation of variable pitch propellers as well as other hydraulically operated systems.

To meet the above requirements, piston engine oils are formulated with typical properties as shown in Table 2. Among the listings, oil viscosity is probably the most important property for all aircraft piston engines. At high operating temperatures, adequate oil viscosity is needed in order to provide an effective hydrodynamic lubrication film for high bearing load. In this case, a higher viscosity index is beneficial as the oil would be more resistant to thinning at the elevated temperatures. Good low-temperature fluidity is also required to allow the aircraft to be cold-started at low ambient temperatures. In this regard, both pour point and viscosity are important parameters in determining the crankability of the engine oil at cold temperatures. Flash point is used to measure the volatility and flammability of the oil and needs to be maintained at an acceptable level. Sulfur compounds in aircraft engine oils are deemed harmful as they can react with moisture and other combustion products to form acids corrosive to aircraft engine bearings and valve guides. An additional measure to determine the level of sulfur corrosivity to yellow metals is the copper strip corrosion test (ASTM D130). Other acidic species that can be corrosive to engine parts and catalytic to oil oxidation are monitored by the total acid number (TAN). To prevent the formation of excessive combustion chamber deposits, ash content must be low, and for most oils the limits are

Aircraft Engine Lubricants, Table 2 Typical properties of aircraft piston engine oils

		Viscosity grade			
Commercial grade	ASTM test method	65	80	100	120
SAE viscosity grade		30	40	50	60
Military grade		1,065	1,080	1,100	1,120
Viscosity (cSt)	D445				
at 100°C		11	15	20	25
at 40°C		95	140	220	280
Viscosity index	D2270	94–115	94–115	94–110	94–120
Pour point (°C)	D97	<−20	<−17	<−17	<−11
Flash point, COC (°C)	D92	230	>240	>250	>250
Ash content (wt.%)	D482	<0.005	<0.005	<0.005	<0.005
Total acid number (mgKOH/g)	D664	<0.1	<0.1	<0.1	<0.1
Sulfur (wt.%)	D2273	<0.5	<0.5	<0.5	<0.5
Copper corrosion, at 100°C	D130	1	1	1	1

Aircraft Engine Lubricants, Table 3 Grades and classes of aircraft turbine engine oils

Specification	Grade/Class	Viscosity at 100°C (cSt.)	Attributes	Typical applications
MIL-PRF-7808	3	3	Low temperature start capability at −54°C or lower	US Air Force fighter jets; commercial aircraft auxiliary power units
	4	4	Improved thermal stability, engine cleanliness	Advanced military aircrafts operating at high temperatures
MIL-PRF-23699	STD	5	Noncorrosion inhibition, balanced stability and load-carrying capability	US Army and Navy turbine engines, commercial and general aviation turbine engines worldwide
	HTS	5	Superior stability and engine cleanliness	
	C/I	5	Corrosion inhibition for corrosive environments	
DOD-PRF-85734		5	EP and enhanced load-carrying properties	Helicopter transmission gearboxes

usually set near zero. In addition to the parameters listed in Table 2, newer generations of ashless dispersant oils are required to have good antiwear and antifoaming characteristics as measured by the ASTM D6709 and D892 test methods, respectively. Recent years have also seen the development of multigrade viscosity oils, including SAE 15W50, 20W50, and 25W60, that meet SAE J1899 specifications. These oils are usually blended with synthetic base oils, along with ashless dispersant additives and shear-stable viscosity index improver additives. They are recommended for use in engines subjected to wide variations of operating temperatures or ambient temperatures due to changes in climate or service region.

Turbine Engine Oils

Most of the aircraft turbine engine oils used today are regulated by two military specifications, MIL-PRF-7808 (1997) and MIL-PRF-23699 (1997). Table 3 lists classification and main applications of the oils, with their basic properties and performance requirements being given in Tables 4 and 5. MIL-PRF-7808 describes oils of two viscosity grades, 3 and 4 cSt (at 100°C), which are designed

Aircraft Engine Lubricants, Table 4 Physical and chemical properties of aircraft turbine oils

	MIL-PRF-7808		MIL-PRF-23699
	Grade 3	Grade 4	STD, C/I and HTS classes
Viscosity (cSt)			
at 205°C (401°F)	–	1.1 min	–
at 100°C (212°F)	3.0 min	4.0 min	4.90–5.40
at 40°C (104°F)	11.5 min	17.0 min	23.0 min
at−40°C (−40°F)	–	–	13,000 max
at−51°C (−60°F)	17,000 max	20,000 max	–
Flash point, COC (°C(°F))	210(410) min	210(410) min	246(475) min
Pour point	–	–	−54°C (−65°F) max
Total acid number, TAN	0.30 max	0.5 max	1.00 max
Evaporation loss, 6.5 h at 204°C (400°F) (wt.%)	–	–	10 max
Evaporation loss, 6.5 h at 205°C (401°F) (wt.%)	30 max	15 max	
Thermal stability and corrosivity at 274°C (525°F)			
Viscosity change (%)		5.0 max	5.0 max
TAN change (mgKOH/g)		6.0 max	6.0 max
Metal weight change (mg/cm^2)		4.0 max	4.0 max

Aircraft Engine Lubricants, Table 5 Requirements of corrosion and oxidation stability for aircraft turbine oils

		Test conditions (ASTM D4636)		Post test oil properties				
Specification	Grade/ class	Temperature	Duration (h)	Viscosity change (%)	TAN change (mgKOH/g)	Weight loss (%)	Sludge content (mg/100 ml oil)	Metal weight change (mg/cm^2)
MIL-PRF-7808	Grade 3	175°C (347°F)	96	−5 to 15	2.0 max	4.0 max	No visible sludge	Must pass for Al, Ag, Bz, Fe, M-50, Mg, and Ti
		200°C (392°F)	96	−5 to 25	4.0 max	4.0 max		
	Grade 4	200°C (392°F)	96	−5 to 18	2.0 max	4.0 max		
		220°C (428°F)	40	−5 to 25	4.0 max	4.0 max		
MIL-PRF-23699	STD & CI	175°C (347°F)	72	−5 to 15	2.0 max	–	50 max	Must pass for Al, Ag, Fe, Mg and Cu
		204°C (400°F)	72	−5 to 25	3.0 max	–	50 max	
		218°C (425°F)	72	Report	Report	–	50 max	Must pass for Al, Fe, Ag, and Ti
	HTS	175°C (347°F)	72	0–10	1.0 max	–	25 max	Must pass for Al, Ag, Fe, Mg and Cu
		204°C (400°F)	72	0–22.5	2.0 max	–	25 max	
		218°C (425°F)	72	Report	Report	–	25 max	Must pass for Al, Fe, Ag, and Ti

for applications that require cold temperature start capability as low as −51°C (−60°F). The more viscous Grade 4 oils have higher thermal stability, thus making them more suitable for advanced aircraft engines that operate at hotter temperatures. MIL-PRF-23699 describes oils of 5 cSt (at 100°C) viscosity grade for use in U.S. Army and Navy turbine applications as well as in most commercial and general aviation turbine engines worldwide. Typically made with neopentyl polyol esters, MIL-PRF-23699 oils have evolved into three major classes: (1) STD or Standard class, (2) HTS or High Thermal Stability class, and (3) C/I or Corrosion Inhibition class. These oils generally exhibit conventional load-carrying capacities and are often referred to as Type II oils. The high load-carrying oils are described in the DOD-PRF-85734 specification (2004). These oils are also of 5 cSt (at 100°C) viscosity grade, but further contain Extreme Pressure (EP) additives, and are typically used in helicopter transmission gearboxes and some gas turbine systems with highly loaded gear sets. The Society of Automotive Engineers (SAE International) recently published Aerospace Specification AS5780 (2005), which sets additional requirements that an oil needs to meet in order to be considered for use. Two classes of oil, both being of 5 cSt (at 100°C) viscosity grade, are described: (1) SPC or Standard Performance Capability class, which closely corresponds to the STD class of the MIL fluids, and (2) HPC or High Performance Capability class, which is similar to the MIL Specification HTS grade. AS5780 is the new standard to define properties and production controls for aero and aero-derived turbine engine lubricants and it is recognized by the Federal Aviation Administration (FAA) in Advisory Circular AC 20-24C (2011) as a component of the overall engine/oil-specific certification process.

Some turbine oils designed for aircraft engines may also be used in nonaviation applications, such as in aero-derived turbine engines used in electrical power generation, petrochemical pipeline compression and delivery, and marine propulsion. These land- and marine-based engines largely retain the essential design and performance characteristics of their aviation siblings, thus requiring similar lubrication and cooling characteristics from the lubricants. When selecting an aircraft turbine oil for a nonaviation application, the properties of the oil must be carefully examined, and the performance verified so as to assure that the oil meets the requirements set by the engine manufacturer.

Cross-References

- ► Gear Lubrication
- ► Lubricant Viscosity
- ► Mineral Oil Base Fluids
- ► Pour Point
- ► Rheology – Viscosity Index
- ► Viscosity Index Additives

References

Aerospace Standard AS5780. *Specification for Aero and Aero-Derived Gas Turbine Engine Lubricants. Revision A. SAE Aerospace, 2005-10*

DOD-PRF-85734A. *Lubricating Oil for Helicopter Transmission System, Synthetic Base*, 29 June 2004

FAA Advisory Circular AC 20-24C. *Approval of Propulsion Fuels and Lubricating Oils. Federal Aviation Administration*, 29 July 2011

A.R. Landsdown, Aviation lubricants, in *Chemistry and Technology of Lubricants*, ed. by R.M. Mortier, S.T. Orszulik (Blackie Academic and Professional, London, 1987)

MIL-PRF-23699F. *Performance Specification for Lubricating Oil, Aircraft Turbine Engine, Synthetic Base, NATO code number O-156*, 21 May 1997

MIL-PRF-7808L. *Performance Specification for Lubricating Oils, Aircraft Turbine Engine, Synthetic Base*, 2 May 1997

H.A. Poitz, R.E. Yungk, Aviation industry, in *Handbook of Lubrication and Tribology, Vol. 1 Application and Maintenance*, ed. by G.E. Totten, 2nd edn. (Taylor & Francis, Boca Raton, 2006)

Aircraft Lubricants

- ► Aviation Turbine Engine Oil Application

Aircraft Piston Engine Oils and Turbine Engine Oils

- ► Aircraft Engine Lubricants

Aircraft Undercarriage Lubricants

- ► Landing Gear Lubricants

ALE - Atomic Layer Epitaxy

- ► Atomic Layer Deposition (ALD)

ALN – Alendronate Sodium

- ► Modified UHMWPE for the Hip Joint (Particle Filled and Reinforced)

ALP – Atomic Layer Processing

► Atomic Layer Deposition (ALD)

Aluminum Oxide

► Materials for Mechanical Seals

Aluminum Rolling

► Chemistry of Rolling Lubricants

Ammonia Nitriding

► Gas Nitriding

Amontons' Laws of Friction

Peter J. Blau
Materials Science and Technology Division,
Oak Ridge National Laboratory,
Oak Ridge, TN, USA

Definition

Dating back to 1699, Amontons' laws of friction assert that there exists a proportionality between friction force and the applied load, and that the friction force is independent of apparent contact area. While well-known in the field of tribology, these so-called "laws" are predated by the work of Leonardo da Vinci and are invalid in many practical situations.

Scientific Applications

Frenchman Guillaume Amontons (1663–1705) conducted research on friction during the late 1600 s and presented his work in a classic paper to the Royal Academy in December of 1699 (Amontons 1699). In the field of tribology, he is best known for two so-called Amontons' laws of friction that derived from one of the conclusions presented in that paper; namely, "that the resistance caused by rubbing only increases or diminishes in proportion to greater or lesser pressure (load) and not according to the greater or lesser extent of the surfaces." In other words:

1. The friction force is directly proportional to the applied load.
2. The friction force is independent of the apparent area of contact.

These relationships appeared to account for the results of early experiments on various pairs of materials and simple machines, leading to an assertion that the proportionality between the friction force and normal force tended toward 1/3. More comprehensive studies of friction using modern instruments have since shown that the frictional relationships put forth by Amontons, and inferred earlier by Leonardo da Vinci (1452–1519), are at best approximations. They do not in general reflect or predict the behavior of interfaces over a wider range of conditions than were originally treated by Amontons. The definition of the friction coefficient (see the article on ► Friction Coefficient) is largely based on the work of da Vinci, Amontons, and Coulomb (Dowson 1998); however, the proportionality between normal and friction forces in a tribosystem is a dimensionless quantity that depends on geometric, mechanical, environmental, thermal, and materials-related characteristics of that tribosystem. Broad generalizations such as Amontons' laws are no longer widely accepted.

Key Applications

Amontons' laws were derived based on macro-scale experiments and observations. While they played an historical role in understanding non-lubricated sliding behavior, they are not generally used in current mechanical design practices where precise measurements of friction forces and friction coefficients, specific to the application, are required. Experimental work or more tribosystem-specific models have largely taken their place.

Cross-References

► Friction Coefficient
► Friction, History of Research

References

G. Amontons, "De la resistance caus"ee dans les machines,' Mémoires de l'Académie Royale, A, Chez Gerard Kuyper, Amsterdam, published in 1706, pp 257–282 (1699)
D. Dowson, *History of Tribology*, 2nd edn. (Wiley, New York, 1998). 768 pp

Amorphous Carbon Coatings

► Diamond-Like Carbon Coatings

Amorphous Tribo-layers

► Self-Mating Metal Articulations in the Hip Joint

Analytical EHL Solution Methods

► Simplified EHL Solution Methods

Analytical Modeling of Rolling Bearings

Pradeep K. Gupta
PKG Inc, Clifton Park, NY, USA

Synonyms

ADORE, advanced dynamics of rolling bearings; Rolling bearing dynamics

Definition

Rolling bearing dynamics modeling represents a generalized solution to differential equations of motion of bearing elements, providing an analytical simulation of real-time dynamic performance of rolling bearings.

Scientific Fundamentals

Fundamental Modeling Elements

The fundamental elements of any triboelement (defined as a mechanical component with significant tribological interactions, rolling bearing being just one example) model are constitutive equations, geometric compatibility, and governing equations. Constitutive equations model material behavior, such as stress/strain behavior of solids and traction slip relations for a lubricant. Geometric compatibility represents certain geometric constraints imposed on component or element behavior. While incompressibility and continuity are conditions that must be met in continuum modeling, imposed external constraints are fundamental to modeling a rolling bearing; a set of preloaded ball bearings, where the motion of the races is constrained in the axial direction, is one example. Governing equations, as the name implies, define the laws that control the motion when the elements are permitted to move. These equations basically relate the applied forces and/or moments to applied displacements. For rolling bearings and various other mechanical components, the governing equations may be classified under three categories (Crandall 1956):

1. Equilibrium: $\sum F = 0$ and $\sum G = 0$
2. Eigenvalue: $AX = \lambda BX$
3. Propagation: $\sum F = m\ddot{x}$; and $\sum G = I\dot{\omega}$

An equilibrium problem states that a compatible displacement field can be obtained by equating all the applied forces and moments to zero. Examples of eigenvalue problems are the computation of natural frequency of vibration or critical speeds of a rotor whereby the system equations are satisfied for non-zero displacements for at least one eigenvalue (λ). Finally, propagation problems equate the sum of applied forces and moments to appropriate accelerations, thereby constituting a real-time dynamic formulation. Thus, an integration of a set of differential equations is necessary to obtain a compatible displacement field. Although a few problems where bearing noise or computation of natural frequency of vibration require an eigenvalue formulation, most applications for rolling bearings are modeled by either an equilibrium or a propagation formulation.

Equilibrium Modeling in Rolling Bearing

Equilibrium modeling in rolling bearing consists of two parts: (1) force equilibrium equations, which determine the relative position of interacting bearing elements, and (2) kinematic considerations, which permit computation of angular velocities of the bearing elements.

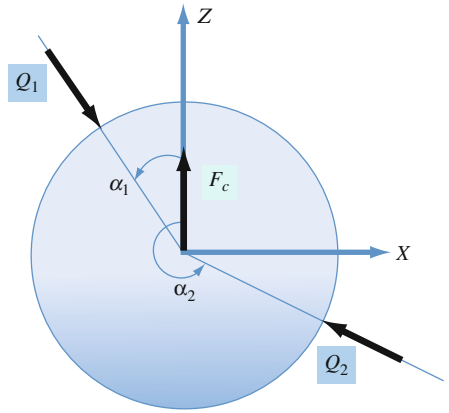
Equilibrium Equations

As shown schematically in Fig. 1, the equilibrium equations of a ball in an angular contact ball bearing are written as:

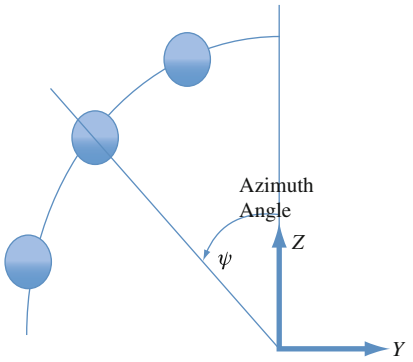
$$\sum_{i=1}^{2} Q_i \sin \alpha_i = 0 \tag{1a}$$

$$\sum_{i=1}^{2} Q_i \cos \alpha_i - F_c = 0 \tag{1b}$$

where the subscript i refers to outer and inner race contacts, Q is the contact load, and α is the contact



Analytical Modeling of Rolling Bearings, Fig. 1 Schematic of ball loads



Analytical Modeling of Rolling Bearings, Fig. 2 Ball angular position and race azimuth

angle. The contact load is a function of the relative axial and radial position of the bearing elements. Thus, the equilibrium equations define the relative axial and radial position of the rolling element relative of the interacting race. Normally, all positions are defined relative to a reference race, generally the outer.

The equilibrium equations for the race are written as:

$$\sum_{i=1}^{n} Q_i \sin \alpha_i + F_x = 0 \tag{2a}$$

$$\sum_{i=1}^{n} Q_i \cos \alpha_i \cos \psi_i + F_r = 0 \tag{2b}$$

where n is the number of rolling elements and F_x and F_r are the applied axial and radial forces, respectively, and ψ is the azimuth angle, which defines the angular position of the rolling element around the bearings, as illustrated in Fig. 2.

Although the above equations are written for ball bearings, they are also applicable to roller bearings, where the contact angles are zero and 180° for the outer and inner races, respectively. Also, depending on the interacting geometry, the applied forces are related to relative positions either by the classical Hertzian point contact solution or the commonly used line contact solutions (Jones 1960; Harris 1966). Thus, the two unknowns in (1a) and (1b) are the axial and radial positions of the rolling element. Similarly, in (2a) and (2b) the two unknowns are axial and radial positions of the race. Due to the nonlinear nature of these equations, a solution is generally obtained by iterative techniques. Thus for a given race position, (1a) and (1b) are solved for each rolling element position. The load solutions are then input into (2a) and (2b) and they are solved for the race position. In the event that the bearing is also subjected to moment loading, the above equations may be further generalized to include applied moments as a function of angular position.

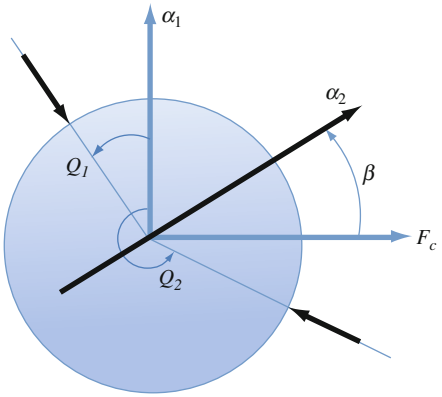
Kinematic Considerations

In addition to computation of relative positions compatible with applied forces, the model also requires computation of rolling element velocities as a function of prescribed race velocities. A kinematic constraint, which is normally imposed on the rolling element to race interaction, requires that at any point in the rolling element to race contact, the surface velocity of the rolling element be equal to that of the race. In other words, there is no slip and the surfaces move in a pure rolling condition. A vector equation for such a constraint may be written as:

$$\vec{u} = (\vec{\Omega} - \vec{\omega}_o) \times \vec{R}_p - \vec{\omega}_b \times \vec{r}_p \tag{3}$$

where $\vec{\Omega}$, $\vec{\omega}_b$, $\vec{\omega}_o$ are, respectively, the race angular velocity, rolling element angular velocity, and rolling element orbital velocity, and $\vec{R}_p$, $\vec{r}_p$; are vectors locating a point in the rolling element to race contact zone relative to the race and rolling element centers, respectively.

For an angular contact ball bearing, as shown schematically in Fig. 3, the ball angular velocity will have two components about the x and z axes, while the race angular velocity and rolling element orbital velocity only have one component along the x axis. Thus, the relative slip $\vec{u}$; will have only one component along the y direction. In the case of pure rolling constraint, this relative slip would be zero. The unknowns in (3) are two components of rolling element angular velocity vector and one component of



Analytical Modeling of Rolling Bearings, Fig. 3 Orientation of ball angular velocity vector

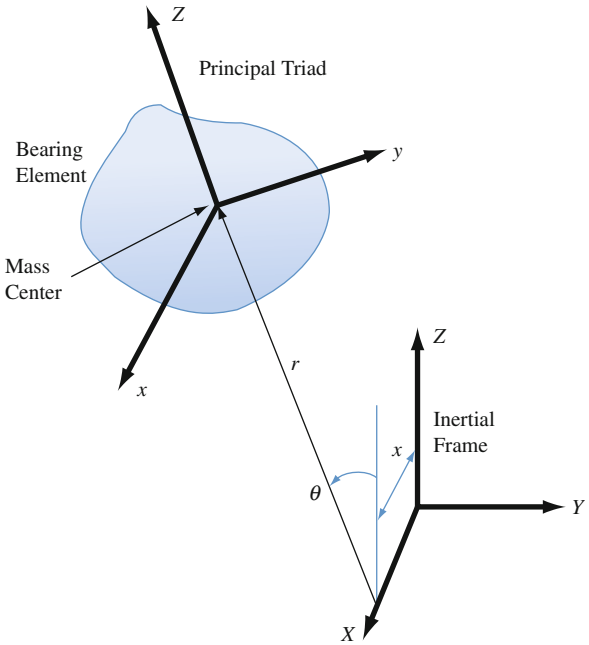
rolling element orbital velocity vector. Application of (3) on both outer and inner race contacts yields two equations. A commonly used *race control* hypothesis (Jones 1960) provides an additional third equation, which is required to obtain a full solution. The hypothesis states that the rolling element angular velocity relative to that of the race about an axis, which is normal to the plane of contact is zero on the race, which provides a smaller friction torque and the pertinent race is said to be the *controlling* race. For computation of friction torque about the axis normal to the plane of contact, a constant friction coefficient is assumed. Although such a constraint is truly an assumption, it has been experimentally found to be valid over a wide range of practical applications.

In lieu of the *race control* hypothesis, an alternate constraint may be to minimize the total frictional energy dissipated in the outer and inner race contacts (Gupta et al. 2002). Under any arbitrary lubrication condition, if the frictional energy due to relative rolling element to race slip is E_1 and E_2 (at the outer and inner race, respectively) and the inclination of the rolling element angular velocity vector is β, as shown in Fig. 3, then the constraint is implemented as:

$$\frac{\partial}{\partial \beta}(E_1 + E_2) = 0 \tag{4}$$

Dynamic bearing performance simulations obtained with high-speed ball bearings under varied lubrication conditions seem to support the above hypothesis (Gupta et al. 2002).

For roller bearings the kinematics are somewhat simpler. The roller angular velocity vector has only one component along the roller axis. Thus, (3) is adequate without any kinematic constraints.



Analytical Modeling of Rolling Bearings, Fig. 4 Cylindrical coordinates for rolling element motion

Dynamic Modeling in Rolling Bearing

In real-time dynamics modeling, the equilibrium equations are replaced by differential equations of motion. In general, the bearing element motion is divided into two parts: motion of the element mass center and rotation of the bearing element about its mass center (Walters 1971; Gupta 1979, 1984). Any rolling bearing is comprised of basically four elements: the rolling elements (ball or rollers), the cage, the outer race, and the inner race. The cage is what separates the rolling elements and prevents them from interacting with each other. The equations of motion for a rolling element are best written in a cylindrical coordinate frame, illustrated in Fig. 4.

$$m\ddot{x} = F_x \tag{5a}$$

$$m\ddot{r} - mr\dot{\theta}^2 = F_r \tag{5b}$$

$$mr\ddot{\theta} + 2m\dot{r}\dot{\theta} = F_\theta \tag{5c}$$

where m is mass of the rolling element, (x, r, θ) are the axial, radial and orbital coordinates, and (F_x, F_r, F_θ) are the components of the applied force vector in the respective directions.

Motion of the cage and the races (both outer and inner) may be best modeled in the Cartesian coordinate frame (X, Y, Z).

$$m\ddot{x} = F_x \tag{6a}$$

$$m\ddot{y} = F_y \tag{6b}$$

$$m\ddot{z} = F_z \tag{6c}$$

where m is the mass of the element being considered and (F_x, F_y, F_z) are components of the applied force vector in the (X, Y, Z) coordinate frame.

In the most generalized fashion, the rotational motion on any bearing element is best modeled by the classical Euler equations of motion (Synge and Griffith 1959) written in a body fixed frame, located along the principal triad (oriented along the three principal axes of the element), as shown in Fig. 4.

$$I_1\dot{\omega}_1 - (I_2 - I_3)\omega_2\omega_3 = G_1 \tag{7a}$$

$$I_2\dot{\omega}_2 - (I_3 - I_1)\omega_3\omega_1 = G_2 \tag{7b}$$

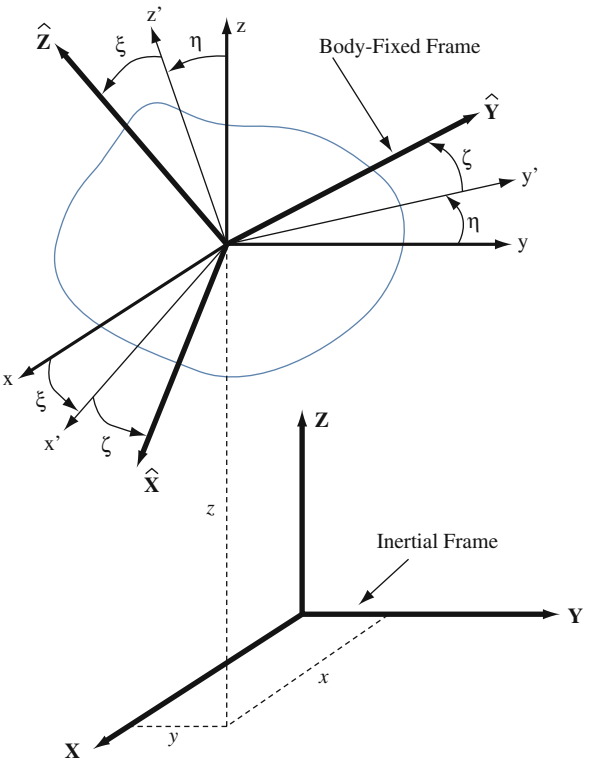
$$I_3\dot{\omega}_3 - (I_1 - I_2)\omega_1\omega_2 = G_3 \tag{7c}$$

where (I_1, I_2, I_3) are the three principal moments of inertia, $(\omega_1, \omega_2, \omega_3)$ are the three components of the angular velocity vector, and (G_1, G_2, G_3) are the three components of the applied moment vector.

Basic Coordinates and Transformation

Implementation of the above equations to develop a model for dynamic performance of a rolling bearing requires definition of certain coordinate frames and procedures to transform a vector between different coordinate frames. For rolling bearings, since the mass center motion is considered in a space fixed coordinate frame and the angular motion is formulated in a body fixed coordinate frame, the two basic coordinates are a space fixed inertial frame and a body frame that is generally located along the three principal axes on the element. In addition to these basic coordinate frames, other coordinates may be defined to facilitate the computation of geometric interaction and applied force and moment vectors.

Based on the existing modeling work (Walters 1971; Gupta 1979, 1984), transformation from one coordinate frame to another is illustrated schematically in Fig. 5, where transformation from inertial to a body-fixed coordinate frame is shown in terms of three rotations: (1) ηI (2) $\xi j'$ (3) $\zeta\hat{k}$ where (I, J, K) and $(\hat{i}, \hat{j}, \hat{k})$ are unit vectors along the corresponding $(X, .Y, Z)$ and $(\hat{x}, \hat{y}, \hat{z})$ axes, and η, ξ, and ζ are the three transformation angles.



Analytical Modeling of Rolling Bearings, Fig. 5 Inertial to body-fixed coordinate transformation

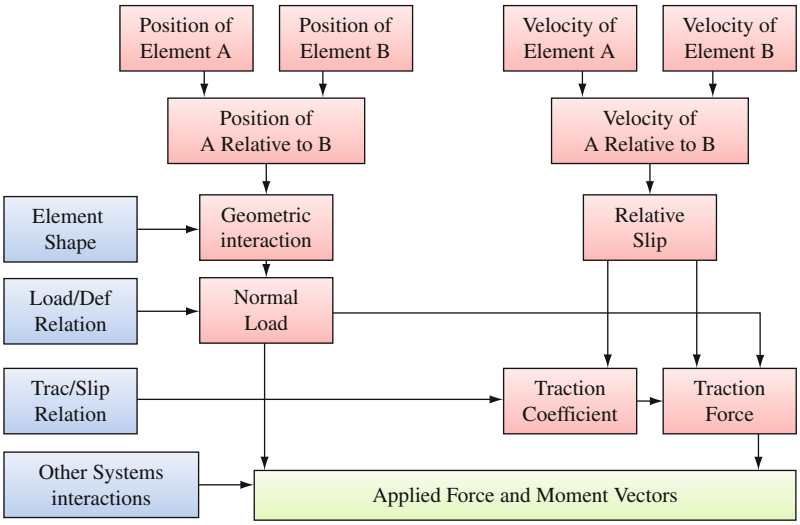
Now using the superscript i for an inertial and b for a body fixed frame, the transformation from inertial to body frame may be written as:

$$r^b = [T_{ib}]r^i \tag{8}$$

where the transformation matrix $[T_{ib}]$ may be written in terms of the three transformation angles, η, ξ, and ζ:

$$[T_{ib}] = [T_{ib}(\eta, \xi, \zeta)] = \begin{bmatrix} \cos\xi\cos\zeta & \cos\eta\sin\zeta + \sin\eta\sin\xi\cos\zeta & \sin\eta\sin\zeta - \cos\eta\sin\xi\cos\zeta \\ -\cos\xi\sin\zeta & \cos\eta\cos\zeta - \sin\eta\sin\xi\sin\zeta & \sin\eta\cos\zeta + \cos\eta\sin\xi\sin\zeta \\ \sin\xi & -\sin\eta\cos\xi & \cos\eta\cos\xi \end{bmatrix} \tag{9}$$

The three transformation angles are quite similar to classical Euler angles and they are found to be most convenient in modeling rolling bearing interactions. Also, the transformation matrix in above (9) is orthogonal. Thus, the inverse of this matrix is simply the transpose.



Analytical Modeling of Rolling Bearings, Fig. 6 Generic architecture of interaction model and applied load computation

In order to complete the formulation, it is essential to establish a relation between the angular velocity vector in the body fixed coordinate frame and the rate of change of the transformation angle. This relation can be shown to be (Gupta 1984)

$$\omega^b = [\mathscr{C}] \begin{Bmatrix} \dot{\eta} \\ \dot{\xi} \\ \dot{\zeta} \end{Bmatrix} \tag{10}$$

where ω^b is the angular velocity vector in body fixed frame and the matrix $\mathscr{C}$ is given by

$$\mathscr{C} = \begin{bmatrix} \cos\xi\cos\zeta & \sin\zeta & 0 \\ -\cos\zeta\sin\xi & \cos\zeta & 0 \\ \sin\zeta & 0 & 1 \end{bmatrix}$$

Although the above matrix is not orthogonal, the inverse is simply shown to be:

$$\mathscr{C}^{-1} = \begin{bmatrix} \frac{\cos\zeta}{\cos\xi} & -\frac{\sin\zeta}{\cos\xi} & 0 \\ \sin\zeta & \cos\zeta & 0 \\ -\tan\xi\cos\zeta & \tan\xi\sin\zeta & 1 \end{bmatrix} \tag{11}$$

which, when combined with (10), yields:

$$\begin{Bmatrix} \dot{\eta} \\ \dot{\xi} \\ \dot{\zeta} \end{Bmatrix} = [\mathscr{C}^{-1}]\omega^b \tag{12}$$

For angular accelerations, the above equation may be differentiated to yield:

$$\begin{Bmatrix} \ddot{\eta} \\ \ddot{\xi} \\ \ddot{\zeta} \end{Bmatrix} = \frac{\partial[\mathscr{C}^{-1}]}{\partial t}\omega^b + [\mathscr{C}^{-1}]\dot{\omega}^b \tag{13}$$

where derivative of the matrix $[\mathscr{C}^{-1}]$ is simply obtained by differentiating (11).

Using the above coordinates and after defining velocities as additional independent variables, the second order differential (5, 6, 7) may be written as a set of 12 first order differential equations in generalized form:

$$\mathbf{x} = \{x, r, \theta, \dot{x}, \dot{r}, \dot{\theta}, \eta, \xi, \zeta, \dot{\eta}, \dot{\xi}, \dot{\zeta}\}^T \tag{14}$$

and the derivative vector

$$\mathbf{y} = \dot{\mathbf{x}} = \{\dot{x}, \dot{r}, \dot{\theta}, \ddot{x}, \ddot{r}, \ddot{\theta}, \dot{\eta}, \dot{\xi}, \dot{\zeta}, \ddot{\eta}, \ddot{\xi}, \ddot{\zeta}\}^T \tag{15}$$

The above vectors are written in terms of polar components used for rolling elements. For cage and the races these components may be replaced by the Cartesian components. This set of 12 first order differential equations for each bearing element is numerically integrated simultaneously to obtain a generalized real-time dynamic simulation of rolling bearing performance. Depending on the applications, the problem may be somewhat simplified by constraining certain degrees of freedom.

Computation of Applied Forces and Moments

The final step in the modeling process is to compute the applied forces and moments from the various position and velocity vectors. A systematic procedure for this computation is outlined in Fig. 6. Basically, from the position vectors of two interacting elements A and B a relative position is computed. Then the geometry of the interacting elements is subtracted to compute geometrical interaction. This geometrical interaction

essentially represents the elastic deformation of the interacting elements that are contacting each other. Now, knowing the elastic deformation, a load deformation model such as a Hertzian point contact or a similar line contact model (Harris 1966) is used to compute the applied loads. Similarly, from the relative velocity of the interacting elements, a slip component, tangential to the plane of contact, is computed. This, along with the contact stress corresponding to the computed normal load, is input into a lubricant traction model to compute a traction coefficient, which is a ratio of the traction (or friction) force to the applied normal load. Thus the applied traction force is computed. Since the position vectors locating the point of contact on each element are also known from the interaction analysis, a cross product of these vectors and the applied load vectors yields the applied moment. These force and moment vectors are input to the equations of motion to complete the modeling process.

Lubricant Traction

Lubricant traction between bearing elements subjected to a rolling/sliding interaction is perhaps the most important parameter that controls overall dynamic performance of a rolling bearing. Therefore, a realistic model to simulate lubricant behavior must be an integral part of the overall bearing dynamics model. Several levels of sophistications are applied in this area.

Hypothetical Model

Based on the early works (Kragelskii 1965), a simple hypothetical model has been promoted to model traction between interacting bearing elements (Gupta 1984). Based on experimentally observed traction/slip behavior, as shown in Fig. 7, a simple algebraic relation may be used to relate the traction coefficient to applied slip velocity:

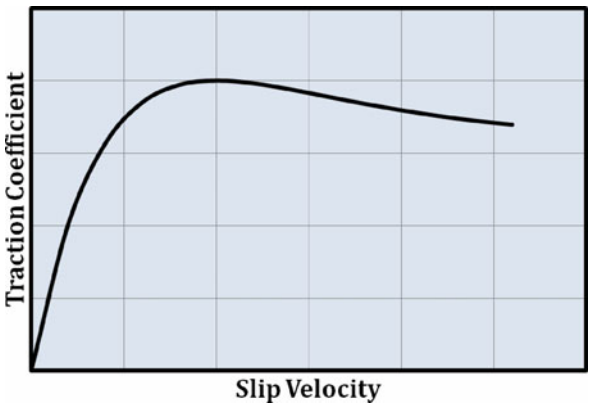
$$\kappa = (A + Bu)\exp(-Cu) + D \tag{16}$$

where κ is the traction coefficient at a slip velocity u, and A, B, C, D are empirical constants, normally derived from experimental traction data.

This simplified model works very well for most solid lubricants, where the traction coefficient is relatively insensitive to rolling velocity and contact pressures. It can also be used for liquid lubricants at a given rolling velocity and contact pressure when the model coefficients are fitted to experimentally observed behavior.

Elastohydrodynamic Models

For most lubricating oils, since the lubricant viscosity is dependent on both pressure and temperature, it is essential to solve the flow equation between elastically deformed surfaces under the applied load. Thus, the model is based on both elasticity of the contacting solids and hydrodynamics of the lubricating fluid, hence it is commonly called the elastohydrodynamic model.



Analytical Modeling of Rolling Bearings, Fig. 7 Hypothetical traction-slip curve

As shown schematically in Fig. 8, an elastohydrodynamic contact between two interacting solids subjected to a normal load basically has a constant film thickness for most of the contact zone under the Hertzian pressure profile. Thus, computation of traction consists of two parts: computation of the film thickness and solution of the combined thermal and mechanical problem in the contact zone. Based on early works, the lubricant hydrodynamic equations have been solved for both line and point contacts, and the solutions are curve fitted to readily usable formulae (Cheng and Sternlicht 1965; Cheng 1970; Hamrock and Dowson 1981). For modeling traction, a somewhat simplified approach is to assume Newtonian behavior of the lubricant in the high-pressure contact region and compute traction by solving the flow equation with prescribed surface velocities and temperatures (Kannel and Walowit 1971). The model is based on the following equations:

Energy equation:

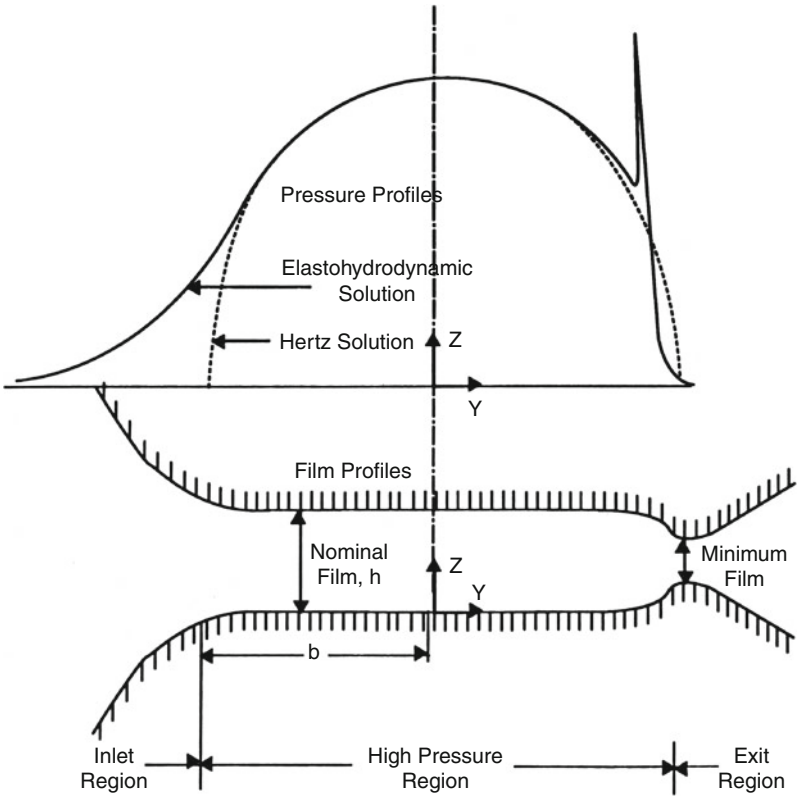
$$k\frac{\partial^2 T}{\partial z^2} = -\tau\dot{s} \tag{17a}$$

Geometric compatibility:

$$\frac{\partial u}{\partial z} = \dot{s}(\tau, p, T) \tag{17b}$$

Constitutive equation:

$$\dot{s}(\tau, p, T) = \frac{\tau}{\mu(p, T)} \tag{17c}$$



Analytical Modeling of Rolling Bearings, Fig. 8 Schematic of an elastohydrodynamic contact

where K is thermal conductivity, T is absolute temperature, τ is shear stress, $\dot{s}$; is shear strain rate, μ is lubricant viscosity, p is pressure, u is the lubricant velocity, and z is the coordinate across the film, as shown in Fig. 8.

The above equations are solved for shear stress and temperature distribution across the film; the shear stress is then integrated over the contact area to compute the total traction force. The primary input to the model is the viscosity-pressure-temperature relation in (17c).

In the Newtonian model, the viscosity varies exponentially both as a function of pressure and temperature. Thus, at very high pressures the Newtonian model yields extremely high viscosities, such that the lubricants tend to behave as a solid rather than a fluid. Under such conditions, a visco-elastic model has been proposed to better simulate lubricant traction (Johnson and Tevaarwerk 1977; Bair and Winer 1979). To simulate such a behavior, a shear stress/strain rate equation is introduced in addition to the viscosity relation used in the Newtonian model:

$$\dot{s} = \frac{1}{G}\frac{\partial \tau}{\partial t} + \frac{\tau_0}{\mu} f\left(\frac{\tau}{\tau_0}\right) \tag{18}$$

Here, G is the shear modulus, τ is the shear stress, and τ_o is defined as a critical shear stress. Two forms of the shear stress functions have been proposed:

$$f\left(\frac{\tau}{\tau_o}\right) = \sinh\left(\frac{\tau}{\tau_o}\right) \tag{19a}$$

$$f\left(\frac{\tau}{\tau_o}\right) = \tanh^{-1}\left(\frac{\tau}{\tau_o}\right) \tag{19b}$$

Computation of shear stress distribution through the lubricant film with the visco-elastic model requires integration of a differential equation, while the Newtonian model may be implemented almost in close form (Gupta 1984). In either model, there are three model coefficients: reference viscosity, pressure-viscosity, and temperature-viscosity coefficients in the Newtonian model, and effective viscosity, shear modulus, and critical shear stress in

the visco-elastic model. Since actual measurement of these constants is extremely difficult, they are generally derived by regression analysis of experimental traction data (Gupta et al. 1981, 1992).

Key Applications

Bearing Life and Stiffness

Bearing fatigue life and stiffness only depend on the normal contact loads at the rolling element to race contacts. Thus, a static equilibrium model is adequate to model these performance parameters. The equilibrium equations are solved by iterative procedures for prescribed applied loads and operating speeds. The solutions provide detailed load distribution over the various rolling elements. These load solutions may be input to semi-empirical life formulae (Harris 1966) to model bearing fatigue life. Since the load deflection relation for a rolling bearing is nonlinear, the bearing stiffness is normally computed from the difference of two solutions obtained in the vicinity of the equilibrium solution with prescribed race displacement.

Rolling Element Skid and Skew

Rolling element skid and roller skew (rotation about the transverse axis) are truly dynamic phenomena. Hence, an integration of the differential equations of motion for the rolling element under the prescribed operating conditions and lubricant behavior is necessary. Also, geometrical imperfections and manufacturing tolerances play a significant role. Therefore, modeling rolling element skid and skew are examples of real applications of the dynamic model outlined above.

Cage Motion and Instability

All interaction forces on the cage are very small in comparison to the applied loads at the rolling element to race contacts. Consequently, they do not affect overall load distribution, bearing fatigue life, and stiffness. However, since these small forces constitute the entire cage loading, they are significant when it comes to modeling overall cage motion and stability. Furthermore, since the rolling elements constantly move back and forth within the cage pockets and collide with the cage in truly dynamic sense, modeling cage motion is yet another application of the real-time dynamic model.

Bearing Heat Generation

Since a realistic simulation of slip between interacting bearing elements requires a solution to the equations of motion, bearing heat generation is another application of the dynamic model. A parametric evaluation of bearing heat generation as a function of bearing geometry and lubricant properties is often valuable in optimizing bearing design and lubricant selection.

Thermal Stability

In most high-speed rolling bearings applications, the bearing heat generation, combined with available cooling systems, alters the temperature distribution through the bearing, which in turn affects internal bearing geometry and, therefore, the loads between interacting bearing elements. These altered loads then feed back to heat generation. It is critical to appreciate this cyclic phenomenon in order to ensure a stable bearing operation. The bearing dynamics models are therefore useful in both optimizing bearing design for a thermally stable operation, as well as prediction of thermal instability as a function of operating environment and bearing design parameters.

Wear

Unlike fatigue, wear failures are commonly associated with some type of instability of bearing element motion. Since most interactions, particularly with the cage, are dynamic in nature, it is often convenient to define a time-averaged wear rate (Gupta 1984):

$$W(T) = \frac{1}{T}\frac{K}{H}\int_0^T Q(t)u(t)dt \qquad (20)$$

where W is time-averaged wear rate over the time interval T, K is wear coefficient, H is the hardness of the material being subjected to wear, Q is the time-dependent load at a given interaction, and u is the sliding velocity as a function of time.

When both Q and u are bounded, the above time-averaged wear rate approaches a steady-state constant value. With realistic values of wear coefficients, this rate can be used to compute wear on pertinent bearing elements and changes in bearing internal geometry. Dynamic simulations may then be repeated using the altered geometry in order to establish an optimal correlation between stability and wear. Under seriously unstable conditions, either the load Q or the slip velocity u, or both, become unbounded, and the time-averaged wear rate, defined above, does not converge to a constant steady-state value. Such conditions lead to a rather catastrophic dynamic bearing failure. The time required to reach such an unstable operation may be defined as the wear life of the bearing.

Cross-References

- ► Elastohydrodynamic Lubrication
- ► Film Thickness Formulas: Line Contacts
- ► Film Thickness Formulas: Point Contacts
- ► Friction/Traction Behavior of EHL
- ► Hertz Theory: Contact of Ellipsoidal Surfaces
- ► Kinematics of Rolling Element Bearings
- ► Radial Bearings
- ► Rolling Element Bearings, History

References

S. Bair, W.O. Winer, A rheological model for EHD contacts based on primary laboratory data. J. Lubrication Technol. ASME Trans. **101**, 258 (1979)

H.S. Cheng, A numerical solution of the elastodydrodynamic film thickness in an elliptical contact. J. Lubrication Technol. ASME Trans. **92**, 155–162 (1970)

H.S. Cheng, B. Sternlicht, A numerical solution for the pressure, temperature, and film thickness between two infinitely long, lubricated rolling and sliding cylinders under heavy loads. J. Basic Eng. ASME Trans. **87**, 695–707 (1965)

S.H. Crandall, *Engineering Analysis* (McGraw Hill, New York, 1956)

P.K. Gupta, Dynamics of rolling element bearings. Parts I to IV. J. Lubrication Technol. ASME Trans. **101**, 293–326 (1979)

P.K. Gupta, *Advanced Dynamics of Rolling Elements* (Springer, Berlin, 1984)

P.K. Gupta, L. Flamand, D. Berthe, M. Godet, On the traction behavior of several lubricants. ASME J. Lubrication Technol. **103**, 55–64 (1981)

P.K. Gupta, H.S. Cheng, N.H. Forster, Viscoelastic effects in MIL-L-7808-type lubricant, Part I: analytical formulation. STLE Tribol. Trans. **35**, 269–274 (1992)

P.K. Gupta, On a kinematic hypothesis for angular contact ball bearings, in *ASTM Symposium on Rolling Element Bearings*, Orlando, Florida, USA, April 2002, pp. 39–47

B.J. Hamrock, D. Dowson, *Ball Bearing Lubrication: The Elastohydrodynamics of Elliptical Contacts* (Wiley, New York, 1981)

T.A. Harris, *Rolling Bearing Analysis* (Wiley, New York, 1966)

A.B. Jones, A general theory for elastically constrained ball and radial roller bearings. J. Basic Eng. ASME Trans. Series D **82**, 309–320 (1960)

K.L. Johnson, J.L. Tevaarwerk, Shear behavior of EHD oil films. Proc. Roy. Soc. Lond. **A356**, 215 (1977)

J.W. Kannel, J.A. Walowit, Simplified analysis for traction between rolling-sliding elastohydrodynamic contacts. J. Lubrication Technol. ASME Trans **93**, 39–46 (1971)

J.L. Synge, B.A. Griffith, *Principles of Mechanics* (McGraw-Hill, New York, 1959)

C.T. Walters, The dynamics of ball bearings. J. Lubrication Technol. ASME Trans. **93**, 1–10 (1971)

Angular Contact Ball Bearings

► Angular Contact Bearings

Angular Contact Bearings

Xiaolan Ai
Timken Technology Center, The Timken Company, Canton, OH, USA

Synonyms

Angular contact ball bearings; Ball bearings for radial-thrust loading; Tapered roller bearings

Definition

Angular contact bearings refer to a type of bearings that support both radial and axial loads. Examples of these bearings are angular contact ball bearings and tapered roller bearings.

Scientific Fundamentals

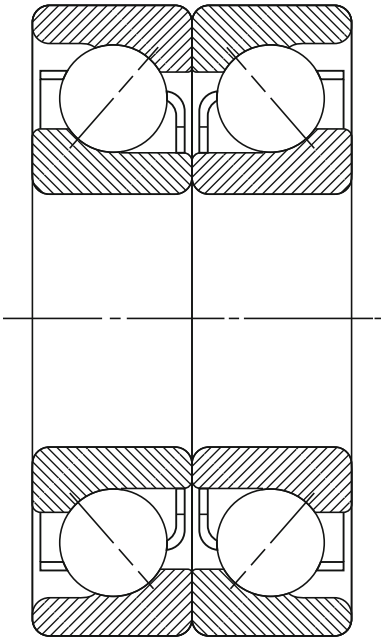
According to loading situations, rolling element bearings are classified into radial bearings, angular contact bearings, and thrust bearings. Angular contact ball bearings and tapered roller bearings are the most common types of angular contact bearing.

Angular Contact Ball Bearings

Angular contact ball bearings (Fig. 1) are designed to carry an appreciable amount of thrust load or a combination of radial and thrust loads. They can be operated at a higher speed than tapered roller bearings. An angular contact ball bearing is comprised of an inner race ring, an outer race ring, and a set of balls interposed between and in rolling contact with the inner and outer race rings. Most standardized angular contact ball bearings further contain a cage or retainer that separates the balls and keeps them spaced apart along the raceways. Angular contact ball bearings can be considered as a variation of deep-groove ball bearings, in that at least one race ring has only one side shoulder. Under zero endplay (lateral movement) conditions, the line that connects the nominal contact between the inner raceway and the ball and the contact between the outer raceway and the ball lies at an angle with respect to the radial plane of the bearing. This angle is referred to as the contact angle. It usually rages from 15° to 45° for commercially available angular contact ball bearings. The point where the line of contact intersects the bearing axis is called the effective bearing center. Under thrust load, the contact angle may increase, exceeding the initial contact angle.

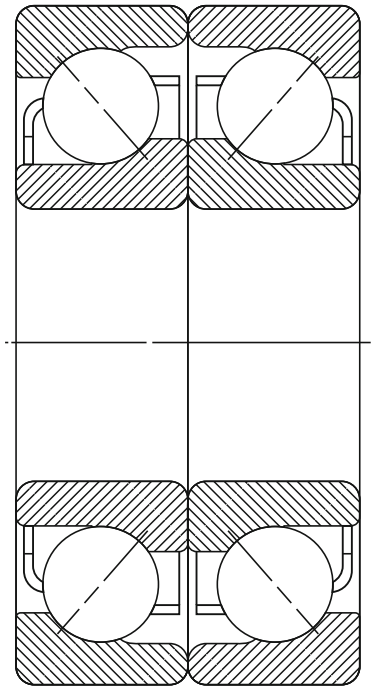
Angular contact ball bearings are commonly used in pairs to provide a greater radial load capacity, or to carry bi-directional thrust loads. Figure 2 shows a back-to-back duplex bearing arrangement. The angular contact lines of

Angular Contact Bearings, Fig. 1 An angular contact ball bearing



Angular Contact Bearings, Fig. 2 Back-to-back duplex bearing arrangement

the two bearings converge on the bearings' back faces. Figure 3 shows a face-to-face duplex bearing arrangement with lines of contact convergent on the bearings' front faces. Since the spread between the effective bearing



Angular Contact Bearings, Fig. 3 Face-to-face duplex bearing arrangement

centers in the back-to-back arrangement is greater than that of the face-to-face arrangement, back-to-back duplex bearings are widely used in applications where a tilting moment exists. Duplex angular contact bearings are preloaded to increase the supporting stiffness and to reduce the displacement due to the applied load as much as possible. The amount of preload can be adjusted by changing the clearance or distance between the two inner rings or between the two outer rings. Preloads are divided into four classes: very light, light, medium, and heavy. As a general rule, grinding machine spindles or machine center spindles require very light to light preloads, and lathe spindles or hypoid gear drives require medium to heavy preloads. Since the preload affects bearing operating torque, temperature, noise, and, more importantly, bearing life, excessive preloading is prohibited. The equation used to determine the amount of preload of duplex bearing set is based on the axial load and displacement relationship given by

$$\delta_a = c_b F_a^{\frac{2}{3}}$$

where δ_a = axial displacement

c_b = displacement constant, determined by bearing internal geometry

F_a = axial preload

Due to angular slippage, or spin motion, between balls and contact surfaces on the inner and outer raceways, angular contact ball bearing experiences some frictional torque, although very small, during operation. Under slow speed or marginal lubrication conditions, the friction torque of the bearing increases in direct proportion to the preload. This torque is often referred to as starting torque or bearing setup torque. Preload can be estimated alternatively by measuring the staring torque. The following equation calculates the bearing starting torque:

$$M_b = c_{sb} F_a^{4/3}$$

where M_b = starting torque, ball bearings

C_{sb} = torque constant for ball bearings

Torque constant C_{sb} varies with the bearing size, internal geometry, and friction conditions at the contacts between the balls and raceways. It is determined specifically for the bearing or bearings under consideration. One should consult with the bearing manufacturer for detailed information.

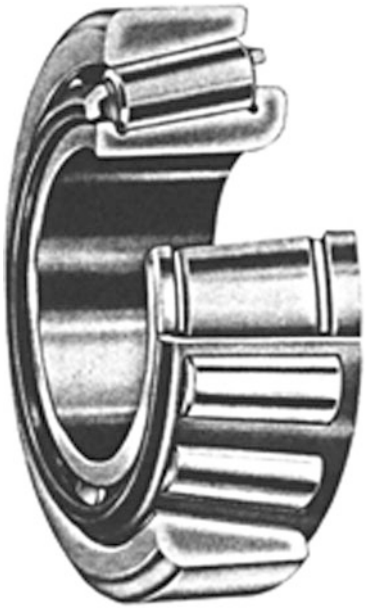
Applications of Angular Contact Ball Bearings

Because of their high rotational precision, high speed capability, low operating torque, low heat generation, and low manufacturing cost, angular contact ball bearings are widely used in machines tool spindles, transmission gear boxes, vehicle wheel ends, electric motors, and household appliances.

Tapered Roller Bearings

A standardized tapered roller has four basic components, as shown in Fig. 4: an inner ring, an outer ring, a set of tapered rollers, and a cage. Some bearing manufacturers refer to the inner ring as the cone and the outer ring as the cup. The cup and the cone have tapered raceways with the tapered rollers guided between them by a very accurately positioned flange know as the rib. The cone, rollers, and cage are assembled together forming the cone assembly. The cup is usually separable from the cone assembly to facilitate machine assembly.

The extensions of the raceways and the tapered roller surfaces are designed to converge at a common point, known as the apex, on the axis of rotation. This provides true rolling motion of rollers on the raceways. Tapered roller bearings are also referred to as angular contact roller bearings. The half included cup angle defines the outer raceway contact angle, and the half included cone angle defines the contact angle for the inner raceway. Under normal operating conditions, the cone, cup, and rollers carry the load while the cage spaces the roller apart and retains them on the cone. The tapered raceways allow the bearing to carry combined radial and thrust loads or thrust load only. The greater the included cup angle, the greater is the ratio of thrust load to radial load capacity. Due to the difference in cup and cone raceway angles, the resultant contact forces between them and the tapered rollers under load are slightly different. The vector difference between the two resultant forces on each roller produces a seating force between the large end of the roller and the cone large rib. The seating force provides positive roller guidance, keeping the rollers aligned. The seating force is a function of the included roller angle and is small compared to the radial and thrust load on the bearing.



Angular Contact Bearings, Fig. 4 A tapered roller bearing

The roller body, cup raceway, and cone raceway are designed to provide full line contacts for maximum load capacity. A tapered roller bearing will carry significantly higher loads than an angular contact ball bearing of comparable size.

Tapered roller bearings are used in pairs like angular contact ball bearings, and are often preloaded to increase bearing rigidity. The relationship between axial load and axial displacement is

$$\delta_a = C_r F_a^{9/10}$$

where $\mathbf{d}_a$ = axial displacement

C_r = displacement constant, determined by bearing internal geometry

F_a = axial preload

The amount of axial displacement of tapered roller bearing is proportional to the axial load to the 0.9 power. The displacement of angular contact ball bearing is proportional to the axial load to the 0.67 power. Thus, there is a greater benefit to increase preload for ball bearings than for tapered roller bearings to maximize the bearing stiffness.

Displacement measurements are routinely used to set up tapered bearings for a targeted preload to ensure healthy operation and long service life. As with angular contact ballbearings, starting torque of a tapered roller bearing can also be used to estimate the preload and thereby to set up the bearing. Under slow rotational speed, the friction torque of a tapered roller bearing is generated primarily by the sliding and spinning motion between the roller ends and cone large rib. Starting torque is particularly useful for production inspections of preassembled bearing packages, as such wheel bearings for automotive application. The starting torque of a tapered roller bearing is sought considering only sliding and spinning friction at the roller end and rib contacts. The following equation provides the relationship between staring torque and axial preload for a tapered roller bearing:

$$M_r = c_{sr} F_a$$

where M_r = starting torque, roller bearings

C_{sr} = torque constant for tapered roller bearings

The torque constant is determined by friction condition at the contacts between the roller ends and the cone rib face, and by internal geometry of the bearing.

Application of Tapered Roller Bearings

The heavy load capacity, along with the ability to carry combined radial and thrust loads makes tapered roller bearings a good choice for a host of applications. Typical applications include wheel and steering axles for automotives and mobile equipment, train axle journals, aircraft landing wheels, transmission gear boxes, industrial pressing machines, steel rolling mills, and machine tool spindles.

Cross-References

► Radial Bearings
► Rolling Bearing Stiffness
► Thrust Bearings

Angular Resolved Light Scattering (ARS)

► Non-Contact Surface Metrology by Means of Light Scattering

Angular-Contact Thrust Bearings

► Thrust Bearings

Ankle Arthroplasty

► Tribology of Ankle Joints

Ankle Joint

► Tribology of Ankle Joints

ANN – Artificial Neural Network

► Drilling Tool Wear Monitoring

Annulus Gears

► Internal Gears

Anti-Adhesion/Stiction Surface Design, Fabrication, and Applications

Xiang-Jun Zhang
State Key Laboratory of Tribology, Tsinghua University, Beijing, People's Republic of China

Definition

Stiction/adhesion is a phenomenon in which two surfaces of a microsystem or of microstructures are adhered together due to various interfacial attractive forces and fail to separate. Thus anti-adhesion/stiction technology

involves a variety of application approaches to overcome the interfacial adhesive forces by structure design and special surface film/coating fabrication.

Scientific Fundamentals

Mechanism

The stiction/adhesion phenomenon of two such surfaces is a consequence of the scaling law of a microsystem: the surface-to-volume ratio scales with the inverse of the microdevice dimensions and interfacial/surface forces of the micro/nanometer scale become dominant. The various interfacial forces include *electrostatic force*, intermolecular ► *van der Waals forces*, and meniscus capillary forces.

With the boom in development of micro/nano systems such as microelectromechanical system (MEMS) and nanoelectromechanical system (NEMS), stiction/adhesion between the substrate (usually silicon based) and the microstructure has become a dominant failure type. A variety of anti-adhesion/stiction approaches have emerged in the related fields of microelectronics, micro/nano tribology, and physical chemistry (Zhao 2003; Zhuang 2005; Delrio et al. 2005).

Stiction occurs when ► *interfacial energy* is higher than the mechanical restoring energy of the microstructure after intimate contact. Actually, in addition to the interfacial adhesion forces, mechanical design of the microstructure, ► *surface roughness* of the substrate and the structure, and environmental conditions (relative humidity or temperature) all play an important role in stiction.

Anti-adhesion/stiction approaches are generally divided into three categories according to the physical causes of stiction: (a) structure design to reduce contact area or avoid intimate contact; (b) avoidance or reduction of meniscus capillary force; (c) surface modification to minimize the ► *surface free energy* of microstructure. The core concept is to avoid or decrease the dominant role of the interfacial adhesion interaction.

Structure Design

It has been proved that stiction may occur when flexible and smooth structures are brought in contact with the substrate (Zhao 2003; Zhuang 2005). Thus, mechanical design, including structure and surface topography, may be an easy way to avoid stiction failure. Firstly, contacting surfaces can be textured by a periodic array of small supporting posts, commonly known as dimples (Lee et al. 2003) to decrease the real contact area, thus reducing the total interfacial attractive forces. Secondly, the microstructure can be designed with a higher structural rigidity to obtain higher restoring energy (e.g., adopting a shorter and thicker beam instead of a flexible thinner one). Additionally, novel structure design can also be considered in real microsystems, such as a "ramps" structure in modern hard disk system, which will be discussed in more detail later.

Avoidance or Reduction of Meniscus Capillary Force

A thin liquid layer between two solid plates can result in significant adhesion forces. If the contact angle between liquid and the solid plates is less than 90° (hydrophilic surfaces), a net attractive force will act between the plates due to the liquid ► *surface tension*, namely capillary force (Zhuang 2005):

$$F_{cap}(d) = \frac{4\gamma r(\cos\theta)^2}{d^2} \quad (1)$$

where γ is surface tension of liquid, θ the contact angle on the solid plates, d the separation distance, and r the Kelvin radius given by Kelvin equation.

For microsystems, a thin liquid layer may result from the fabrication process or environment humidity during use. According to equation (1), effective anti-adhesion/stiction measures may include roughening or hydrophobic modification of the surfaces. The ► *surface roughness* can increase the minimum separation d and decrease the real contact area. The hydrophobic surface can increase the contact angle and prevent a water layer forming between the two surfaces. In engineering, many microdevices are designed to be encapsulated into an inert gas cell to avoid water condensation.

Surface Modification

It has been found that micro/nano-scale stiction is closely related to ► surface free energy of the used materials. The material surfaces with higher ► surface free energy have higher adhesion force and thereby higher tendency for stiction. Effective approaches may be to fabricate surface film/coatings with low surface-energy materials, such as a *self-assembly monolayer (SAM)* (Zhao 2003; Maboudian et al. 2000) or *diamond-like carbon (DLC)* coatings (Li et al. 2008).

The term *SAM* denotes a single layer of ordered special molecules adsorbed onto a substrate due to bondings between the surface and molecular head groups. The head groups of the SAM chemisorb on the substrate. The chains realize a hydrophobic property to reduce the surface energy significantly. Several classes of organic films have been explored. These include alkyi- and

perfluoroalkytrichlorosilane SAMs, dichlorosilane- and alkene-based molecular films. Among these, the most widely used is the octadecyltrichlorosilane (OTS)-based film.

Key Applications

Anti-Adhesion/Stiction Approaches in Hard Disk Drives

In the context of hard disk drives, stiction refers to the tendency of a read/write head to stick to the platter. Stiction most likely occurs as a result of properties of the platters (smoothness and magnetic forces) as well as other forces including interfacial ►van der Waals force, the cohesion of lubricant thin film to form a meniscus, and so on.

Firstly, breakdown of lubricant thin film may cause the read/write head to stick to the platter. Today, with the increase of scanning speed and the decrease of flight height, the much tighter head-flatter space causes much higher internal operating temperatures, which often lead to an accelerated breakdown of the surface lubricants. On another hand, the several-nanometer head-flatter separation may also cause cohesion of the lubricant thin film, which may form a liquid meniscus of 10 nm (Izumisawa et al. 2002), much higher than the flight height. The resulting capillary force may cause severe stiction of the read/write head. Therefore, a suitable selection of lubricant (including ►surface free energy, molecular weight, and end groups) becomes more important for anti-stiction.

Another stiction case is that the read/write head "sticks" to the parking zone after shut-down of the computer. In 1995, a contact start/stop (CSS) technology was adopted in IBM to improve the anti-stiction performance. The CSS area was designed as a landing zone on the platter, usually near its inner diameter where no data is stored. The landing zone was made by producing an array of smooth nanometer-scale "bumps," thus vastly decreasing the real contact area and, thus, the interfacial adhesion forces.

Today, modern hard drives have mostly solved this kind of stiction problem by a head unloading technology, which includes a "ramps" structure to "unload" the head from the disk surface during shut-down, similar to the concept of a "airport" for the head. These ramps ensure the head is not touching the platter when it stops or takes off to achieve to a high speed. This approach not only prevents stiction but also prevents abrasion from kicking up microscopic particulates that may later contaminate the drive mechanism.

Cross-References

►Self-Assembled Monolayers
►Surface Free Energy
►Surface Tension
►Van der Waals Forces

References

F.W. Delrio et al., The role of van der Waals forces in adhesion of micromachined surfaces. Nat. Mater. **4**(8), 629–634 (2005)
S. Izumisawa et al., Stability analysis of ultra-thin lubricant films with chain-end functional groups. Tribol. Lett. **12**(1), 75–81 (2002)
C.C. Lee et al., Method on surface roughness modification to alleviate stiction of microstructures. J. Vac. Sci. Technol. B **21**(4), 1505–1510 (2003)
X. Li et al., DLC films as anti-stiction coatings for MEMS, in *Proceedings of the SPIE – The International Society for Optical Engineering*, Shenyang, China, **7133**, 713347–713352 (2008)
R. Maboudian et al., Self-assembled monolayers as anti-stiction coatings for MEMS: characteristics and recent developments. Sens. Actuators **82**, 219–223 (2000)
Y.P. Zhao, Stiction and anti-stiction in MEMS and NEMS. Acta Mech. Sinca **19**(l), 1–10 (2003)
Y.X. Zhuang, On the stiction of MEMS materials. Tribol. Lett. **19**(2), 111–118 (2005)

Antifriction

►Friction and Lubrication in Electrical Contacts

Antifriction Materials and Composites

►Solid Lubricants, Polymer-Based Self-Lubricating Materials

Antiwear Additives

►Ashless Phosphate Esters

Antiwear Chemistry

►Tribochemistry of Antiwear (AW) Additives

Application of Mixed EHL in Gears

► Mixed EHL in Gears

Applications of DLC in Magnetic Recording

Charanjit Singh Bhatia[1], Ehsan Rismani-Yazdi[2], Sujeet K. Sinha[2], Aaron James Danner[1]
[1]ECE Department, National University of Singapore, Singapore, Singapore
[2]Department of Mechanical Engineering, National University of Singapore, Singapore, Singapore

Synonyms

Tribology of diamond-like carbon in magnetic recording

Definitions

Magnetic storage and magnetic recording are terms from engineering, referring to the storage of data on a magnetized medium using different patterns of magnetization.

Diamond-like carbon, or DLC, is an amorphous carbon containing significant amounts of sp^3 bonding to prevent wear and corrosion.

Scientific Fundamentals

Introduction

Presently, hard disk drives (HDDs) are the most common devices for mass storage of information. Hard drives are magnetic storage devices in which data is recorded as magnetized bits on the surface of a thin layer of ferromagnetic materials like cobalt or its alloys. A read-write head flying a few nanometers above the surface of this magnetic media is responsible for writing and recovering the recorded data on the disk. Surfaces of the magnetic media and the head must always be protected against corrosion and against mechanical damages (wear and tear) whenever intermittent contact occurs between the head and the disk. One way to protect the head/media interface is to coat these surfaces with an extremely thin, continuous, and hard material that is also chemically inert and atomically dense and also with an overcoat of a lubricant layer. This lubricant acts as an additional corrosion barrier and also lowers the friction between the head and the disk.

Amorphous carbon films can meet approximately all of these requirements. Many different types of carbon coatings have been used as protective layers on the surfaces of magnetic disks and heads. Since its invention, the performance of the HDD has continuously improved, and its recording density has been dramatically increasing every year (see Fig. 1). Data storage density increases exponentially with the decrease in spacing between the magnetic media and the read-write elements on the head. This requires a continuous reduction in the thickness of the carbon coatings while preserving their unique mechanical and chemical properties. This has necessitated improvements in the preliminary procedures, compositions, and deposition methods of these protective carbon films.

This article will first give a short introduction to carbon and its different allotropes. This is followed by a discussion of amorphous carbon films and different methods of deposition. Finally, key applications of DLC films in HDDs will be highlighted

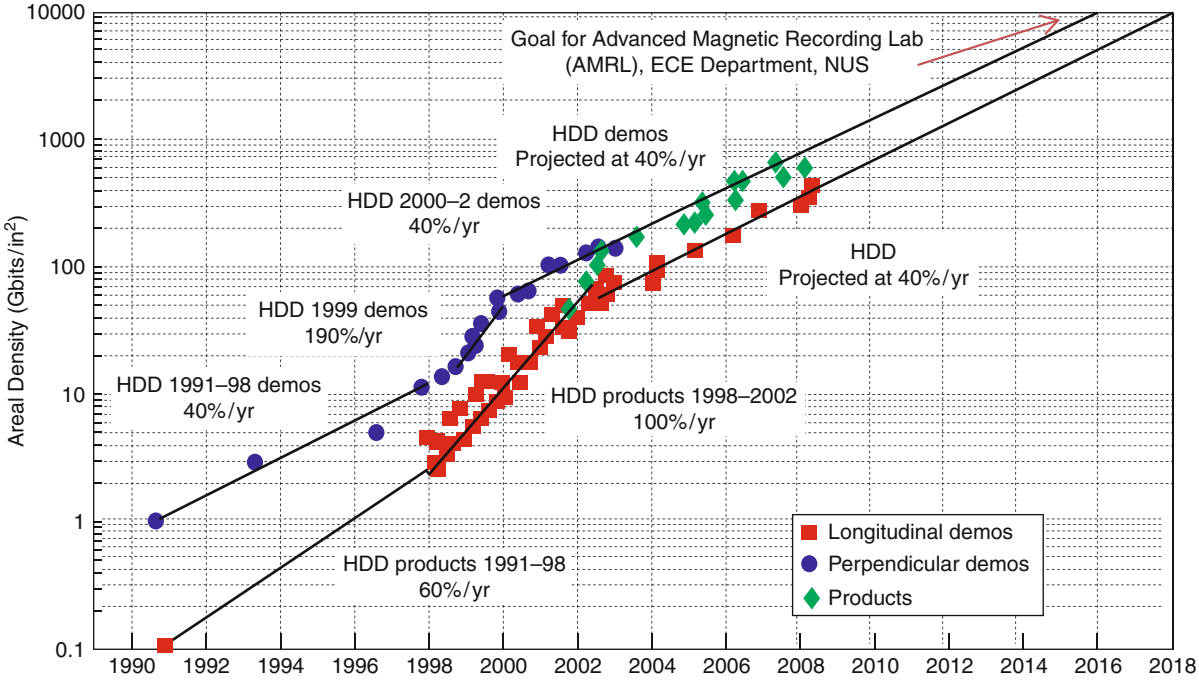
Diamond-like Carbon (Definition and Fundamentals)

Carbon is a very versatile element that can be found in crystalline (diamond and graphite) and amorphous forms. Carbon atoms are able to form three different types of hybrid orbitals (sp^1, sp^2, and sp^3) and therefore can bond to each other or to other atoms in different ways.

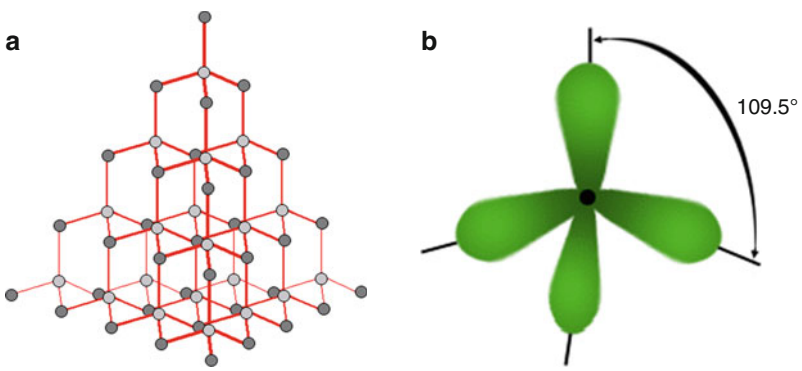
Allotropes of Carbon

Carbon atoms in different forms may have different configurations or allotropes. In cubic diamonds, each carbon atom forms four identical sp^3 hybridized orbitals arranged tetrahedrally around the atom (Fig. 2b). Carbon atoms are connected to their neighbors by means of overlapping of these similar sp^3 orbitals, forming a three-dimensional lattice of carbon atoms bonded together with very strong covalent bonds (Fig. 2a). This makes diamonds the hardest material that we know of.

In ordered graphite, one s and two p orbitals are mixed to form three identical sp^2 hybrid orbitals with a triangular planar configuration (Fig. 3b). Each atom is connected to its three neighbors (σ bond) to form a hexagonal network of carbon atoms (graphene plane) in extended layers (Fig. 3a). The bonds between the carbons within the layer are stronger than those in diamond.



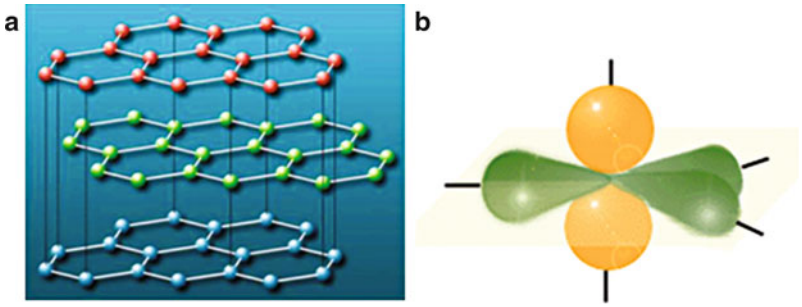
Applications of DLC in Magnetic Recording, Fig. 1 Increasing recording density of hard disk drives over time



Applications of DLC in Magnetic Recording, Fig. 2 (**a**) Carbon atoms making a giant macromolecular array (lattice) in diamond. (**b**) sp^3 hybridization of a carbon atom

The un-hybridized remaining electron that is perpendicular to this plane forms a π bond with an orbital similar to the below or above layer. This bond is not very strong. Therefore, graphene layers can slip on each other, and this makes graphite a soft and lubricious material. In addition, this electron is de-localized (mobile) and allows graphite to conduct electricity.

Amorphous carbon (a-C) is an allotrope of carbon that is not crystalline, and the atoms are connected to each other with a disordered combination of sp^2 and sp^3 bonds to form clusters or rings or a chain network configuration (Fig. 4). Based on the various fractions of sp^2 and sp^3 bonds (C-C sp^3, C-C sp^2, C-H sp^3, etc.), this material may exhibit completely different chemical and physical behaviors.

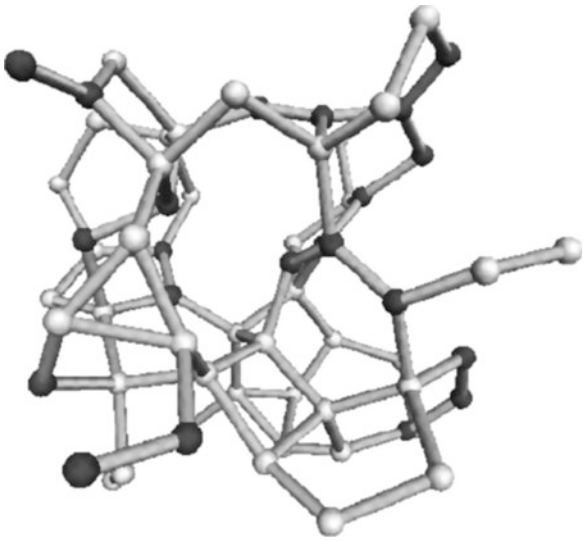


Applications of DLC in Magnetic Recording, Fig. 3 (**a**) Configuration of carbon atoms in graphite. (**b**) sp^2 hybridization of a carbon atom

An amorphous carbon containing a higher portion of C-C sp^3 bonds like those that connect carbon atoms in the lattice of diamond can show some desired mechanical and chemical properties such as high hardness and Young's modulus, high smoothness, chemical inertness, and high atomic density. Because of the similarity of its properties to those of diamond, this type of amorphous carbon is called diamond-like carbon, or DLC.

Unlike diamond, which has a high growth temperature, DLC thin films can be manufactured near room temperature. The first hard, amorphous carbon (a-C) films were grown by Aisenberg, Chabot, and Holland in the 1970s using a plasma deposition method. This work was continued by Koidle and coworkers in the 1980s to produce hydrogenated amorphous carbons (a-C:H). Meanwhile, the first cathodic arc systems were invented and a graphite source was used to grow a new generation of DLC films with tetrahedrally bonded carbon atoms that contain no hydrogen or nitrogen. This type of DLC is called tetrahedral amorphous carbon, or ta-C (Bhatia et al. 1998; Donnet and Erdemir 2008; Anders 2009).

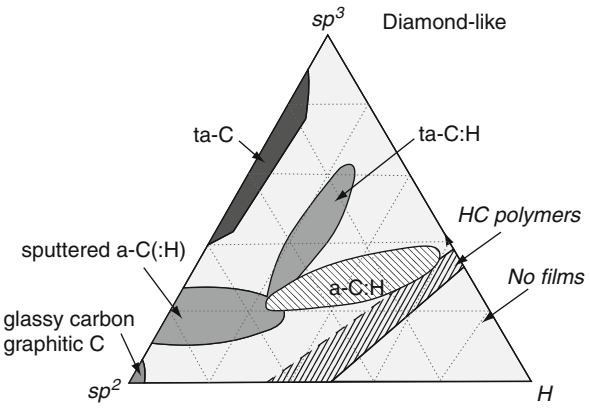
Properties of DLC films can be carefully adjusted by controlling the distribution of the carbon sp^1, sp^2, and sp^3 hybridizations and the amount of hydrogen and nitrogen atoms in the film, as well as the deposition method. Because of their versatile properties, DLC films have greater importance and are used in many different scientific and engineering fields such as tribology, magnetic data storage, food storage, auto industries, optics, and bioengineering (Casiraghi et al. 2007). Raman spectroscopy (Casiraghi et al. 2005), x-ray photoelectron spectroscopy (XPS), electron energy loss spectroscopy (EELS), and electron and neutron diffraction are some of the most important non-destructive characterization methods (Pharr et al. 1996) to determine sp^3 content of a DLC film.



Applications of DLC in Magnetic Recording, Fig. 4 Structure of C-C bonds in amorphous carbon

Different Types of DLCs and Their Properties

DLC coatings can be categorized in three major groups: hydrogenated (hydrogenated amorphous carbon or a-C:H), nitrogenated (a-C:N or CN_x), and hydrogen-free (amorphous carbon a-C and tetrahedral amorphous carbon ta-C) films. Other types of DLCs that are available are doped or alloyed DLCs.



Applications of DLC in Magnetic Recording, Fig. 5 Ternary phase diagram of the amorphous carbon hydrogen system (Robertson 2002)

Hydrogenated DLC

Compositions of carbon films (without nitrogen) are usually shown on a ternary phase diagram of hydrogen, sp^2, and sp^3 hybridization, as shown in Fig. 5 (Robertson 2002).

Amorphous carbons such as soot, chars, and glassy carbons with a disordered graphitic ordering lie in the lower left-hand corner of this diagram. The bottom right-hand corner of the triangle (with H content greater than 70%) corresponds to hydrocarbon molecules or soft polymers rather than hard solid continuous covalent networks. Although they have more than 70% sp^3, most of these positions are terminated by H-C bonds.

Amorphous carbon with intermediate hydrogen content (20–40%) is called amorphous hydrogenated carbon (a-C:H). This group lies at the center of the ternary diagram, and because it has a considerable amount of C-C sp^3 it is categorized as DLC. a-C:H films can be grown by reactive sputtering of carbon in the presence of hydrogen gas or by plasma-enhanced chemical vapor deposition (Ferrari 2004).

Keeping the hydrogen content in a range of 25–30% and using deposition methods with a high-density plasma source, the C-C sp^3 bonds can be increased up to 70%. Because of the existence of the H and a covalent network of tetrahedrally (sp^3) connected carbon atoms. These films are called tetrahedral hydrogenated amorphous carbon or ta-C:H and these films show better mechanical properties compared to a-C:H films.

Hydrogen-Free DLC

By simple sputtering of graphite, an amorphous carbon (a-C) film with very low C-C sp^3 content will form that contains almost no hydrogen. This material is soft but still dense enough and can be used as a corrosion-resistant layer.

A hydrogen-free amorphous carbon with very high fraction of C-C sp^3 (tetrahedral) bonds is called tetrahedral amorphous carbon (ta-C). These films can be grown by deposition methods that contain a highly ionized plasma of almost equal ion energy such as filtered cathodic vacuum arc (FCVA), mass-selected ion beam (MSIB), or pulsed laser deposition (PLD) (Robertson 2002; Anders 2009). The ta-C films may have the maximum amount of sp^3 (up to 85%), which is directly responsible for the better mechanical and chemical properties of the films. The maximum fraction of sp^3 can be obtained at an ion energy of about 100 eV (Pharr et al. 1996).

Nitrogenated Amorphous Carbon

Incorporation of nitrogen into amorphous carbon films has a profound effect on their properties. Introducing N_2 gas into a-C deposited by sputtering of graphite at a temperature of 200°C causes the formation of a nanostructured film in which graphite planes are strongly crosslinked together (Hultman et al. 2001). Although the sp^2 content in this film is still high, because of the increase in the disorder, the mechanical properties of the film (elastic recovery and hardness) are improved.

Deposition of carbon by methods that produce highly ionized plasma like FCVA, MSIB, and PLD in a very low pressure N_2 atmosphere will form tetrahedral nitrogenated amorphous carbon (ta-C:N_x) films with high sp^3 content of 80–90%. Because of their unique mechanical and chemical properties, nitrogenated amorphous carbon films are being used as protective layers against corrosion and wear on the surface of magnetic hard disks (Li et al. 2002).

Doped or Alloyed DLC

Although high compressive internal stress has a key role in the formation of sp^3 bonds and the DLC's mechanical strength, this stress makes DLC films (especially thicker ones) very brittle. This causes rapid propagation of cracks and finally failure of the coating under high loads and impacts. In addition, formation of DLC films with a high compressive stress needs excellent adhesion between the film and substrate, otherwise by increasing the film

thickness or by applying external loads the film will be delaminated or peel off the surface.

Alloying the carbon films with suitable metals or nonmetal elements may improve some of these limitations and enhance the mechanical properties of carbon films. For example, alloying of films with Si facilitates the formation of sp^3 bonding in the absence of high compressive stress. Hydrogenated carbon films tend to be graphitized at high temperatures. Alloying these films by Si not only increases their thermal stability but also improves their frictional behavior at higher relative humidity. Metal alloying (Cu, Ni, or Ag) is also used to improve the hardness and toughness of the carbon films in a similar manner (Robertson 2002).

DLC Thin Film Deposition Methods

Amorphous carbon films can be grown by completely different physical or chemical deposition methods. The deposition method directly affects the properties and behaviors of the deposited films. Among the many available methods, sputtering, filtered cathodic vacuum arc, and plasma-enhanced chemical deposition methods are briefly explained.

Sputtering

Sputtering is one of the basic methods for deposition of thin films of different materials such as metals, ceramics, and carbon. In this method, the surface of the source material, which is called the target, is bombarded by energetic particles such as Ar^+ ions and, if the plasma (Ar^+) particles that are colliding with the target have enough energy to overcome the binding energy of the target surface, the surface atoms of the target (for the case of a-C deposition, the target is high-purity graphite) will be ejected into the environment of the sputtering chamber. These particles, which may contain a combination of atoms and ions, will deposit (condense) on the surface of the substrate. The energy and the angle of the bombarding plasma are two key parameters to control during the sputtering process.

In sputter deposition of carbon, most of the produced particles (atoms) are neutral and this mixture has a low energy of about 5 eV. For this reason, the deposited particles cannot penetrate into the outer surface of the substrate and overcome the coating nucleation obstacles; therefore, the particles landing on the substrate surface move toward each other to form separate islands to lower the surface energy. This process causes formation of a discontinuous and rough overcoat for thicknesses less than 2 nm. However, the islands will later merge together, forming a continuous overcoat. Because of the low energy of the sputtered particle this sputtered a-C film has very low content of sp^3 and is not very hard (Robertson 2001).

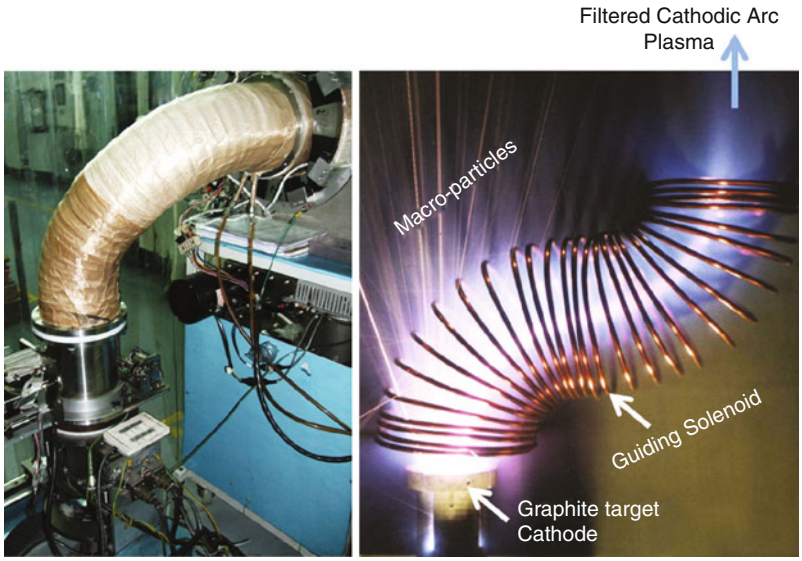
By sputtering graphite in an atmosphere of argon and hydrogen plasma, a-C:H will form and similarly formation of a-CN_x will be carried out in a nitrogen-argon plasma environment.

Plasma-Enhanced Vapor Chemical Deposition

Vapor chemical deposition is a deposition method in which a solid thin film is formed from a chemical reaction between one or more chemical precursors in the vapor or gas phase. The required energy to initiate the reaction (to chemically decompose the reactive gases) is provided by heating the whole reaction chamber up to 600°C. This is the main disadvantage of this method to deposit carbon films as this high temperature leads to graphitization and overall deterioration of film properties (Cuomo et al. 1991). Thin films can be deposited by other CVD methods working at a lower temperature. Plasma-enhanced chemical vapor deposition (PECVD) is one of the CVD methods in which the chemical reaction is initiated by the introduction of plasma of the reacting gases to the precursor gases. The energetic electrons in the plasma supply extra energy to the gas molecules in the reactor for initialization of the process and formation of ions and radicals, which in turn lowers the deposition temperature.

For deposition of a-C:H, the reactor chamber is filled with hydrocarbon precursors such as C_2H_2 or CH_4. Plasma is formed by decomposition of the hydrocarbon molecules between two parallel plates (electrodes) when an RF voltage is applied to the plates. The sample is bombarded by the resulting positive ions and radicals, and the coating is formed by the adsorption of these ionized particles on the surface and then chemically bonded to the other atoms on the sample surface.

In this method, the fraction of ionized particles in the plasma is not much (<10%). This cannot lead to the formation of carbon films with high a fraction of C-C sp^3 bonds and a good adhesion to the substrate. These drawbacks can be overcome by exploiting high plasma density sources such as electron cyclotron wave resonance (ECWR), inductively coupled plasma (ICP), plasma beam source (PBS), or helicon sources that create more ionized plasma (Casiraghi et al. 2007). Some of these methods that work at low pressure (PBS or ECWR) are able to produce DLC films with a high fraction of



Applications of DLC in Magnetic Recording, Fig. 6 Configuration of a filtered cathodic vacuum arc with an S-shaped filter

C-C tetrahedral bonds (ta-C:H) with enhanced properties compared to a-C:H film.

Filtered Cathodic Vacuum Arc (FCVA)

Filtered cathodic vacuum arc is an established technique for producing metallurgical and hard coatings. It was first noted in 1970s that carbon films deposited by this method exhibited diamond-like properties (Anders 2009). The plasma is produced by striking a low-voltage, high-current electric arc on the surface of the graphite source (target or cathode) in a vacuum chamber. The produced plasma is highly ionized but contains many macro particles and clusters that can be considered as defects in most cases where a very thin carbon overcoat is needed.

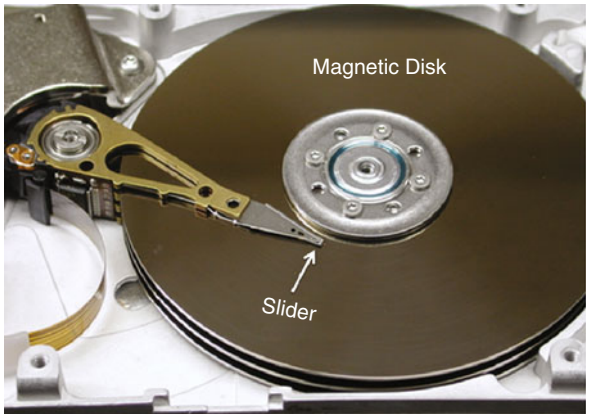
The quality of the plasma can be improved dramatically by using magnetic macroparticle filters. These filters are magnetic ducts bent at 90° angles from the arc source and the plasma is guided out of the duct. The magnetic field causes electrons to follow the filter shape. Because of the electrostatic potential of the electrons, the positive carbon ions follow the same path. However, clusters and unionized particles, because of their higher momentum or having no electric charge, are not affected by this field and continue their motion in a straight line and get trapped on the duct surface. Although most of these particles are filtered by this method, some may return to the plasma by bouncing back from the walls. These bounced particles can be removed by adding another 90° bent filter in series (in plane or out of plane) with the first filter to make a so-called "S-bent filter" (Fig. 6).

This filtered plasma is approximately fully ionized with a kinetic energy of about 20 eV (depending on the arc current) and can form a highly sp^3 bonded carbon (ta-C) film when deposited on the surface. The energy of the ions leaving the S-bent filter can be increased by accelerating the ions to the substrate surface. This is carried out by applying a negative bias voltage to the substrate. The properties of the cathodic arc carbon films can be controlled by varying the ion energy, which is easily achieved through substrate DC or pulsed biasing.

In the FCVA deposition method, if reactive gases such as nitrogen are introduced during the plasma formation process, because of interaction between the ion flux and the gas compounds, films such as ta-CNx can be deposited.

Magnetic Recording (Hard Disk Drives)

Magnetic storage (recording) is basically defined as recording information by local magnetization of a thin film of ferromagnetic material in opposing directions by an external magnetic field. This ferromagnetic material is known as the "magnetic media" and can be coated on flexible polymeric substrates (magnetic tapes) or rigid



Applications of DLC in Magnetic Recording, Fig. 7
Configurations of disk and slider

glass or aluminum plates (hard disks). The recording magnetic field is made by the write head, which is usually moving with respect to the magnetic media. This head then can read the recorded information by measuring the variations of the magnetic field above the film surface.

Data bits are written along tracks. The areal density is defined as the number of bits along one inch of a track (bits per inch, or BPI) multiplied by the number of tracks per inch (TPI) of the disk. The ratio of TPI and BPI is called bit aspect ratio. Each bit is written by magnetization of a number of ferromagnetic grains of the media.

Hard disk drives are the most common method to store data. In hard disks, the magnetic medium consists of a thin film Co-based alloy (Co-Cr-Pt) deposited (sputtered) on the surface of a glass or aluminum alloy substrate. The disk rotates at a speed of approximately 5,400–15,000 rpm. The read-write head consists of many layers of thin film. The head is not in contact with the media and flies in a close proximity over the rotating disk. An Al_2O_3/TiC (AlTiC) ceramic slider supports the magnetic sensing elements. AlTiC is a hard composite ceramic consisting of about 70% by weight Al_2O_3 and 30% by weight TiC. This ceramic part is the main bearing surface when the head is flying above the media or when any crash occurs between head and disk (Fig. 7).

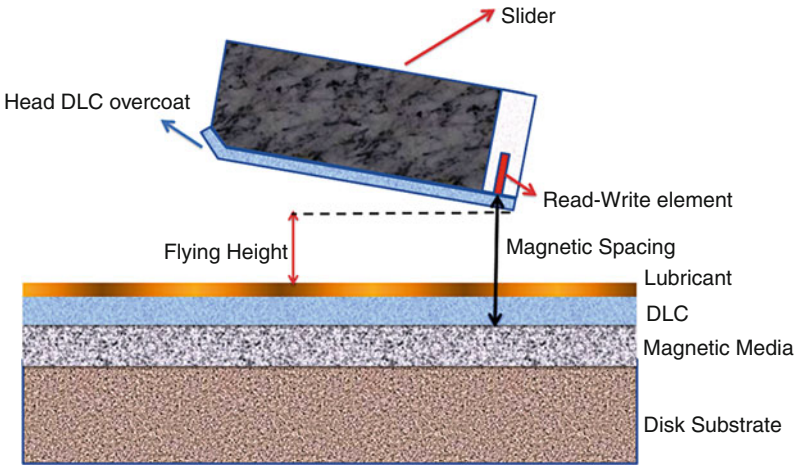
To prevent the magnetic medium and the head read-write elements (magnetic sensor and poles) from oxidation and wear, the surfaces of the media and head are coated with a thin carbon (DLC) overcoat (Fig. 8).

The top surface of the DLC layer is covered by very thin layer of liquid lubricant. This lubricant should have a low vapor pressure, good viscosity, and high thermal stability. It also should be chemically inert and should not react with the other materials in the head-media interface. Perfluoropolyethers (PFPEs), such as ZDOL and Fomblin, have all of these superior properties and are extensively used as lubricants on the surfaces of the magnetic media in disk drives. The combination of the lubricant and hard carbon coating can lower the friction and wear between the media and head when they are in contact for any reason. The lubricant can also prevent media corrosion.

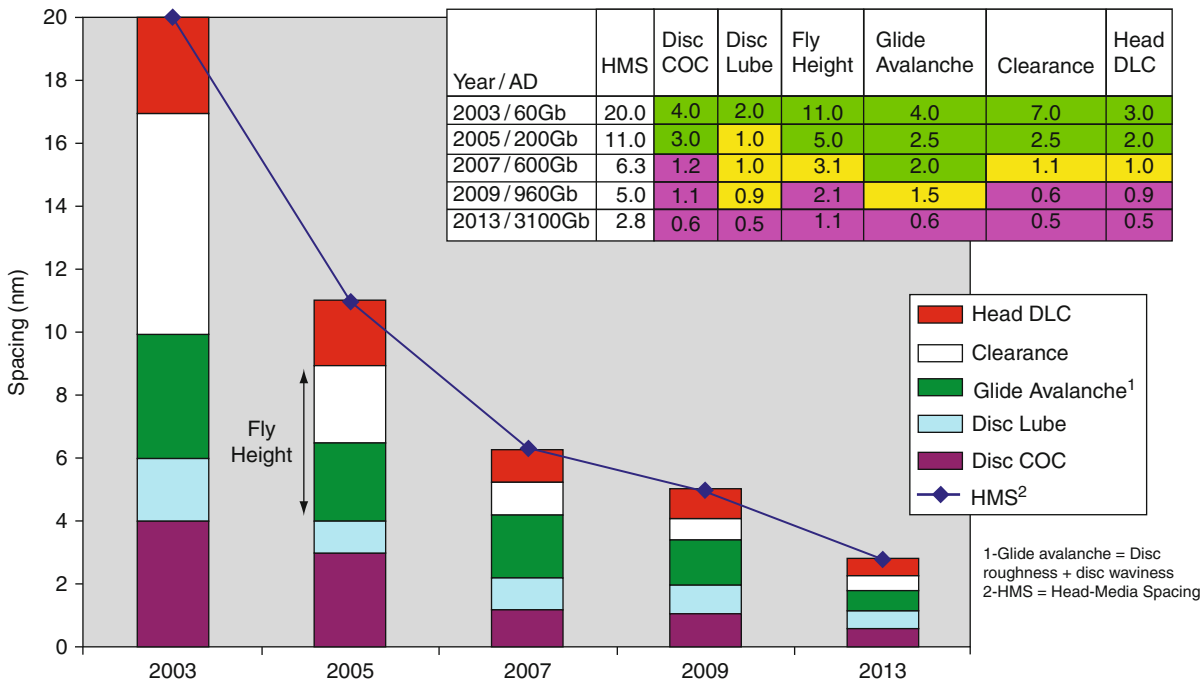
Since its invention by IBM in 1956, the areal density of hard disk drives has continuously increased at a rate of about 40% every year, from 0.002 Mbits/in^2 to 1 Tbits/in^2 today. This has been accomplished by reducing the size of bits (reduction of the bit aspect ratio) and magnetic grains, improving the signal processing methods to recover the recorded data with higher signal-to-noise ratio, and using magnetic materials with higher coercivities (Ferrari 2004).

The vertical distance between the read-write elements of the head and magnetic media is known as the magnetic spacing, which consists of the flying height of the head, the thicknesses of the carbon overcoat and lubricant, and the surface roughnesses of the head and media. According to the Wallace equation (Wallace 1951), bit size and therefore data storage density increases exponentially with a decrease in the magnetic spacing. Magnetic spacing can be reduced by decreasing the fly height as well as the thickness of the carbon films (see Fig. 9). In traditional hard disks with an areal density of 10 Gbit/in^2, the magnetic spacing was approximately 25 nm, which allowed a protective carbon layer on the head and media not thicker than 5 nm. In modern hard disks of 1 Tbit/in^2 areal density, DLC coating should not be thicker than □2 nm. To protect the magnetic materials against wear and corrosion, these films should be atomically smooth, continuous, dense, and hard.

To meet the requirements of new generations of magnetic disks, an ideal overcoat should be able to protect the magnetic surface against corrosion and wear while its thickness is still lower than 2 nm. This requires the coating to cover the surface completely without any defect or pinholes. The film should also be very dense, to act as a diffusion barrier. This material should be compatible with the lubricant and provide good adhesion



Applications of DLC in Magnetic Recording, Fig. 8 Schematic diagram showing the head-media interface



Year / AD	HMS	Disc COC	Disc Lube	Fly Height	Glide Avalanche	Clearance	Head DLC
2003 / 60Gb	20.0	4.0	2.0	11.0	4.0	7.0	3.0
2005 / 200Gb	11.0	3.0	1.0	5.0	2.5	2.5	2.0
2007 / 600Gb	6.3	1.2	1.0	3.1	2.0	1.1	1.0
2009 / 960Gb	5.0	1.1	0.9	2.1	1.5	0.6	0.9
2013 / 3100Gb	2.8	0.6	0.5	1.1	0.6	0.5	0.5

Applications of DLC in Magnetic Recording, Fig. 9 Evolution of various components of the head-media spacing with time key applications of DLC in magnetic recording

sites for the lubricant molecules. This coating should be as smooth as possible and should be extremely hard, tough, and have a low coefficient of friction to protect the surface against wear.

Because of their unique properties, DLC films have been the only candidate for coating the surface of the magnetic layers in hard disk drives. Initially, carbon films (a-C with a high fraction of sp^2 bonds) were sputtered on

the surface of magnetic disks to act as a corrosion barrier. After this, amorphous hydrogenated carbon films with higher sp^3 content were used to protect the media against wear and mechanical damages as well as corrosion. These films were generally deposited by magnetron sputtering.

The head's read-write elements (magnetic sensor and poles) are made of mechanically soft magnetic materials and have a greater wear rate than the surrounding hard AlTiC ceramic. This difference in the wear rates results in a recession of the magnetic poles. This process is called pole tip recession (PTR) and increases the magnetic spacing between the active elements of the head and the media. PTR can be reduced by coating the head with an ultrathin hard DLC film.

Head overcoat not only prevents wear and corrosion of the head elements but also is necessary as a barrier against lubricant degradation. Frictional effects can lead to a decomposition of the PFPE molecule and emission of fluorine. Fluorine can react with a-C:H films and form HF and CF_2O. The use of uncoated Al_2O_3/TiC sliders leads to the rapid decomposition of PFPE molecules. HF and CF_2O react with aluminum oxide in the slider, which is a Lewis acid former to form AlF_3. Lewis acid catalytic reactions on the AIF_3 surface can rapidly decompose PFPE lubricant molecules. Coating sliders with DLC films can be a remedy to reduce the intensity of lubricant decomposition primarily due to frictional actions (Bhatia et al. 1999).

Nitrogenated carbon (CN_x) films are better materials for use in disk protective overcoats due to their desirable wear performance, low coefficient of friction, toughness, and strong adhesion on a disk surface. CN_x films have better compatibility with existing lubricants. The above-mentioned catalytic reactions are prevented due to less hydrogen evolution from the CN_x overcoat, resulting in better tribological performance for the CN_x film as compared to that of the CH_x film (Chen et al. 2000).

Although hydrogenated and nitrogenated carbon overcoats have been widely used, the durability of these films (mostly sputtered films) falls off markedly at thicknesses below 5 nm. Moreover, these films are not able to form continuous films thinner than 2 nm. This required researchers to find an alternative overcoat material such as hydrogen-free tetrahedral amorphous carbon films produced by FCVA. These films are harder than amorphous carbon films produced by other deposition methods; indeed, they exhibit a hardness approaching that of diamond. Because of their toughness and low coefficient of friction, these FCVA ta-C films have been well studied for head/disk interface applications. In many cases, the tribological properties of the head-disk interface are greatly improved by applying these films to the slider and/or media surface.

A number of carbon overcoat film properties vary as a function of ion energy or bias voltage (Pharr et al. 1996). Many research groups have shown that the highest sp^3 fraction (more than 85%) can be obtained at ion energy of □100 eV. The sp^3 content of a film can be measured by methods like EELS, XPS, and Raman spectroscopy. This is much higher than the sp^3 content of sputtered-deposited hydrogenated carbon films, which have a sp^3 content of 15–30%. Other properties of the film such as mass density, hardness, elastic modulus, and internal compressive stress directly correlate with the sp^3 content. Hardness and elastic modulus of thin films are measured by the nanoindentation method. At □100 eV the hardness and elastic modulus of the FCVA ta-C films are at a maximum. At this bias voltage the hardness of the film ranged up to □80 GPa, compared with □20 GPa for hydrogenated, sputtered carbon films. The elastic modulus of these films is also remarkably high, indicating that they are also very tough.

Film stress can be measured by deposition of a film on a thin wafer and measuring the curvature of the wafer before and after deposition. The highest stress also occurs at ion energy of □100 eV. High stress can impair the adhesion of the FCVA films to the substrate. In order to improve the adhesion of the carbon film to the substrate, materials such as Si, Cr, Ti, W, and Mo can be used as underlayers. In addition, a multilayer film with alternative hard and soft layers of carbon and other materials, or carbon layers with different sp^3/sp^2 fractions can decrease the stress levels in the film. This facilitates deposition of thicker films with better durability (Anders et al. 1997).

The thermal stability of hard carbon films is very important in most applications. Deposition of carbon overcoats at higher temperatures leads to graphitization and deterioration of the film properties (Robertson 2002; Erdemir and Donnet 2006). While hydrogenated carbon films begin to soften at temperatures as low as 200°C, the FCVA ta-C films show high hardness and sp^3 content even at temperatures up to 400°C.

Surface roughness of the FCVA films produced at □100 eV is very small (□0.1 nm), and finally, the mass density of FCVA ta-C films measured by Rutherford backscattering spectroscopy (RBS) or EELS is at a maximum 3 g/cm^3, which is close to that of diamond (3.52 g/cm^3). This makes this coating act as a barrier against corrosion on the surfaces of magnetic disks and heads (Pharr et al. 1996).

Recently, FCVA ta-C overcoats thinner than 1 nm have been applied on the surfaces of magnetic films. It has been found that these thin films can still provide adequate protection on the substrate and show good corrosion and oxidation resistance. In addition, new attempts have been made to investigate surface modification of the magnetic medium of hard disks by energetic C^+ ion bombardment under controlled FCVA conditions to produce surface modification and prevent corrosion resistance and give good tribological protection. This could assist in the production of "overcoat-free" magnetic media and heads that exhibit increased mechanical strength and good oxidation resistance. If this can be achieved, then the magnetic spacing (□2 nm) needed for 10 Tbit/in^2 will be possible. This should enable hard disk drive technology to go beyond 2017.

Cross-References

► Lubricants for Rigid Magnetic Disks
► Tribochemistry of Computer Disks

References

A. Anders, *Cathodic Arcs: From Fractal Spots to Energetic Condensation* (Springer, New York, 2009)
S. Anders, D.L. Callahan et al., Multilayers of amorphous carbon prepared by cathodic arc deposition. Surf. Coat. Technol. **94–95**, 189–194 (1997)
C.S. Bhatia, S. Anders et al., Ultra-thin overcoats for the head/disk interface tribology. J. Tribol. **120**(4), 795–799 (1998)
C.S. Bhatia, F. Walton et al., Tribo-chemistry at the head/disk interface. Magn. IEEE Trans. **35**(2), 910–915 (1999)
C. Casiraghi, A.C. Ferrari et al., Raman spectroscopy of hydrogenated amorphous carbons. Phys. Rev. B Condens. Matter Mater. Phys. **72**(8), 1–14 (2005)
C. Casiraghi, J. Robertson et al., Diamond-like carbon for data and beer storage. Mater. Today **10**(1–2), 44–53 (2007)
C.Y. Chen, W. Fong et al., Initiation of lubricant catalytic decomposition by hydrogen evolution from contact sliding on CHx and CNx overcoats. Tribol. Lett. **8**(1), 25–34 (2000)
J.J. Cuomo, D.L. Pappas et al., Vapor deposition processes for amorphous carbon films with sp_3 fractions approaching diamond. J. Appl. Phys. **70**(3), 1706–1711 (1991)
C. Donnet, A. Erdemir, *Tribology of Diamond-Like Carbon Films: Fundamentals and Applications* (Springer, New York, 2008)
A. Erdemir, C. Donnet, Tribology of diamond-like carbon films: recent progress and future prospects. J. Phys. D: Appl. Phys. **39**(18), R311 (2006)
A.C. Ferrari, Diamond-like carbon for magnetic storage disks. Surf. Coat. Technol. **180–181**, 190–206 (2004)
L. Hultman, S. Stafström et al., Cross-linked nano-onions of carbon nitride in the solid phase: existence of a novel C48N12 aza-fullerene. Phys. Rev. Lett. **87**(22), 225503 (2001)
D.J. Li, M.U. Guruz, et al., Ultrathin CN_x overcoats for 1 Tb/in2 hard disk drive systems. Appl. Phys. Lett. 81, 1113 (2002)
G.M. Pharr, D.L. Callahan et al., Hardness, elastic modulus, and structure of very hard carbon films produced by cathodic-arc deposition with substrate pulse biasing. Appl. Phys. Lett. **68**(6), 779–781 (1996)
J. Robertson, Ultrathin carbon coatings for magnetic storage technology. Thin Solid Films **383**(1–2), 81–88 (2001)
J. Robertson, Diamond-like amorphous carbon. Mater. Sci. Eng. R Rep. **37**(4–6) (2002)
R.L. Wallace, The reproduction of magnetically recorded signal. Bell Syst. Technol. J. **30**, 1145–1173 (1951)

Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

Copper has very good electrical and thermal conductivity. Its low cost compared with silver make it a possible arcing contact material in some designs. Cu and Cu alloys such as brass (CuZn) or bronze (CuSn) are commonly used for conductors but generally do not perform as arcing contact tips because of contact resistance and welding concerns. Cu/W composites are the most common arcing contacts for high current switching up to several thousand amperes in vacuum, oil and SF6 switches. Cu/Mo and Cu/WC can perform similarly but less popular nowadays.

Scientific Fundamentals

When copper is used as an arcing contact, the high temperature arc can melt the contact severely and cause strong erosion and contact welding. If this occurs in air, copper is also oxidized to form high-resistance copper oxides. Therefore, for copper to be suitable, the design must have high contact force to reduce contact resistance and wiping action to mechanically wear out the oxide products. We see this used in some knife switch designs and rotary switches where wipe action is readily available. Still, this is limited to very rudimentary designs with low current of less than 50 A and low residential voltages.

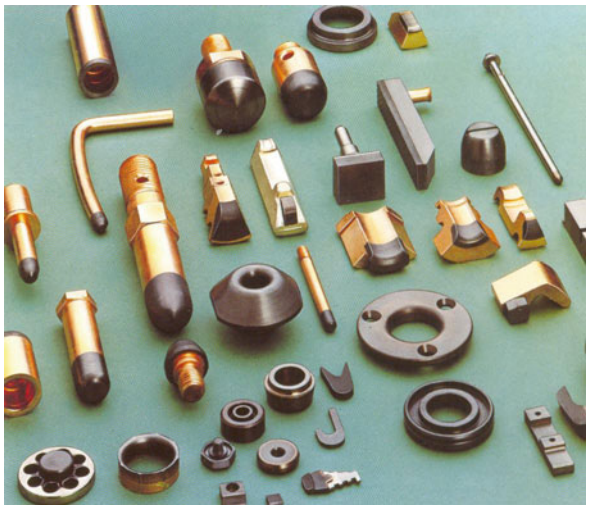
The good electrical conductivity of copper is better used in composite contacts. Large power switching contacts with 10–50 wt% copper can be found in composites

Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites, Table 1 Characteristics of some common Cu/W as electrical contacts (Doduco Data Book 1977)

Material	Characteristics	Processing	Applications
Cu/W	At the common composition of 75–90% W, it is hard material with a strong tungsten to tungsten particle inter-connected skeleton that is resistant to arc erosion. This is used in vacuum interrupters, SF6 circuit breakers, and oil circuit breakers for power generation, transmission, and distribution networks	Discrete contacts are made by a powder metal process to near net shape. Metallurgical modifications of tungsten particle size, distribution, and process temperature can alter electrical performance	Power circuit breakers (SF6, oil, and compressed air switches) with high short circuit currents: >40 kA CuW75 to CuW90 <40 kA CuW65 to CuW75 Vacuum load interrupters 36 kV 3 kA Vacuum contactors Vacuum switch disconnect Transformer tap changers

Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites, Table 2 Physical properties of copper-tungsten powder metal contact (Doduco Data Book 1977)

Material	Cu	Refractory	Typical density (g/cm^3)	Typical hardness	Electric conductivity ($m/\Omega\text{-}mm^2$)
WCu10	90	W 10	16.9	B105	14–18
WCu20	80	W 20	15.4	B100	16–20
WCu25	75	W 25	14.8	B95	17–21
WCu30	70	W 30	14.1	B90	18–22
WCu40	60	W 40	13.0	B80	20–24
WCu50	50	W 50	12.0	B65	22–26
CuWC50	50	50	11.0	B95	26

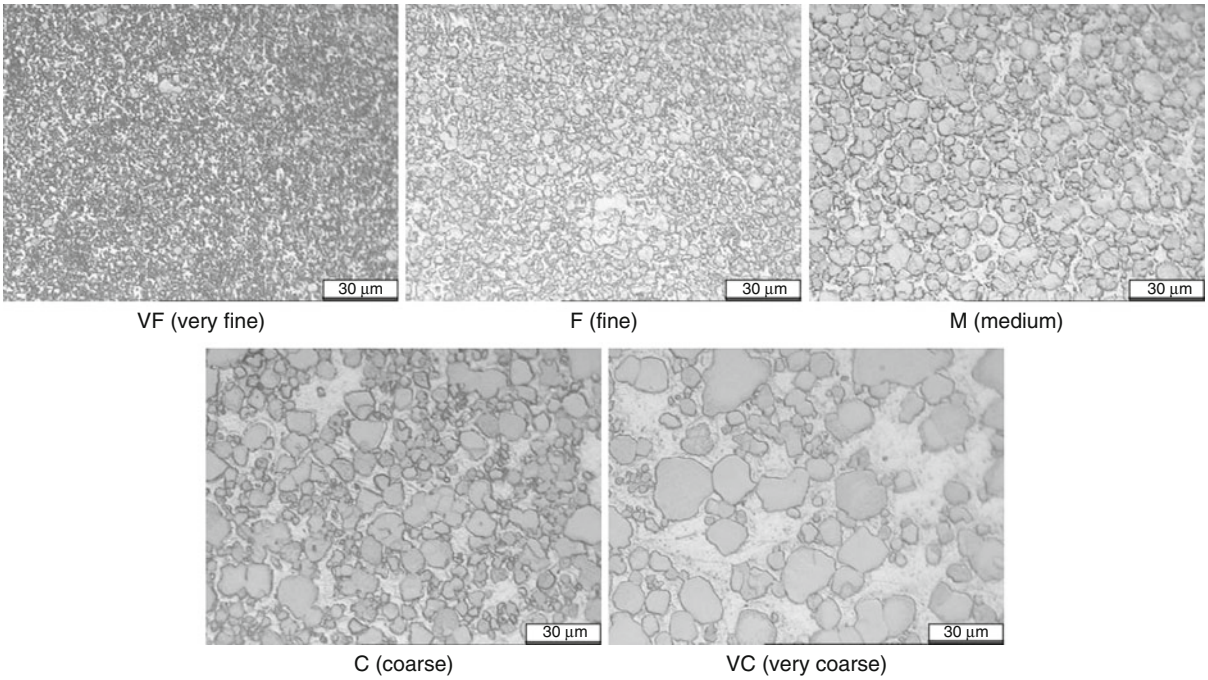


Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites, Fig. 1 Examples of CuW contacts

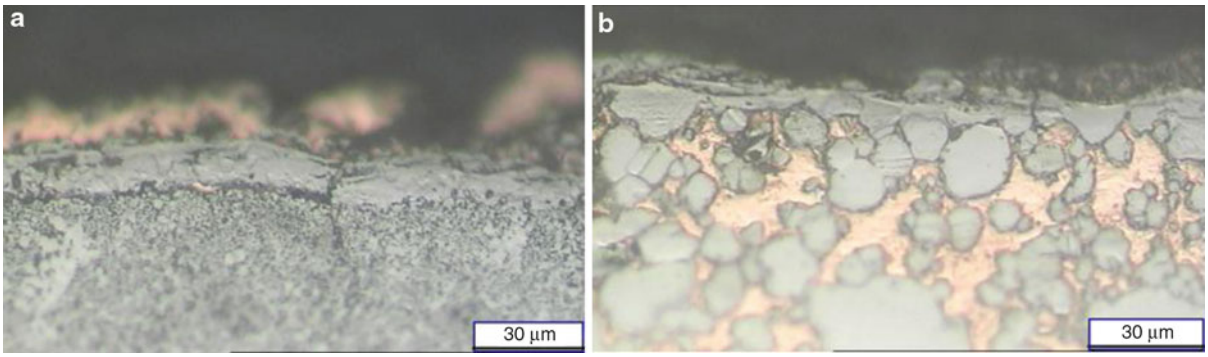
with W. The refractory components have high melting temperature to reduce probability of welding. Copper oxide formation is avoided in vacuum interrupters, oil breakers, and sulfur hexafluoride gas-sealed interrupters. If copper oxide is to form, its contact resistance can be compensated with supplemental low resistance main contacts.

Copper/tungsten composites are manufactured by powder metal processes. In the process, copper powder and tungsten metal powder in the particles several microns in size are thoroughly mixed, granulated, and pressed into green contact shapes. Depending on the process parameters of composition, particle size and other sintering aids, this is sintered at high temperatures of 800–1,500°C in a reducing atmosphere to full density, or can be sintered and then infiltrated with additional molten copper to full density. The sintering develops a strong tungsten skeleton to maintain the contact shape when heated by the high arc temperature of over 3,000°C (Tables 1 and 2).

Leung et al. (2003) reported the microstructure effect on reignition and welding properties of copper-tungsten electric contact (Fig. 1). Figures 2 and 3 are photographs of the metallurgical modifications that can affect contact performance.



Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites, Fig. 2 Optical micrographs of tungsten particle sizes from 1 to 15 µm. All are of the same composition of 22Cu/78W (by weight) (Leung et al. 2003)



Arcing Contact Materials, Cu, Cu Alloy, and Cu-Refractory Composites, Fig. 3 Cross section of 22Cu/78W weld-tested sample: (**a**) very fine tungsten and (**b**) very coarse tungsten showing difference in modified tungsten thickness

Key Applications

Circuit breakers, vacuum interrupters

References

Doduco Data Book, Manual for Engineers and Businessmen, 2nd edn. (Trade brochure, Pforzheim, 1977)

C. Leung, E. Streicher, D. Fitzgerald, D. Ilich, Microstructure effect on reignition and welding properties of copper-tungsten electric contact, *49th IEEE Holm Conference on Electric Contacts* (Washington, DC, 2003)

Arcing Contact Materials, Gold Alloys

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

Gold and gold alloys are used as arcing contact tips to make and break low-power electric circuits, preferably in

the sub-watt range. This material is necessary in miniature technology and high-reliability switches and relays where space limits design parameters. Proper material selection is necessary to maintain low contact resistance and reliable switching performance.

Scientific Fundamentals

Arcing is defined as a spark or gaseous discharge that occurs between a pair of electric contacts with the result of material melting and erosion because of the heat generation. A minimum open circuit voltage and a minimum current are necessary to create the necessary electric field intensity and sufficient Joule heating to have a sustainable discharge and current flow. This minimum arc voltage and minimum arc current are material dependent and influenced by surface conditions of cleanliness. Gold and its alloys have minimum arc voltages and currents around 12–14 V and .4 A. In many electronic circuits with less system voltages and currents, it is possible to have inductive and capacitive elements that create transient arcing conditions.

This arcing process generates localized arc heat to cause melting and contact erosion. This can result in contact welding as well as oxidation. Contact welding is undesirable because these miniature mechanical devices have low breaking force to open the contact tips. Contact oxidation is also undesirable because of low available contact force to break any insulating oxide films. Gold and its alloys are suitable contact materials in such low energy arcing applications and are used in low power electronic and telecommunication switches and relays. In miniature signal relays for switching less than minimum arcs or "dry circuit" conditions, the reliability of gold contacts is even better. Transistor switches are replacing some of the miniature relays, but where cost, safety, and reliability are considered, in many instances the latter is still preferred.

Gold and gold alloys are chosen as electric contacts for several reasons (Tables 1 and 2):

(a) Low environmental corrosion that keeps the contact surface clean and low surface resistance.
(b) Low contact force in the milli-newtons range is sufficient to maintain reliable connectivity.
(c) High melting temperature to be resistant to contact welding.

Arcing Contact Materials, Gold Alloys, Table 1 Physical Properties of gold and gold alloys (Doduco Data Book 1977)

Material	Density (g/cm^3)	Melting point melting range (°C)	Electric conductivity (m/Ω-mm^2)
Au (99.9)	19.3	1,063	44
Au Ag8	18.1	1,058	16.3
Au Ag20	16.4	1,036–1,040	10
Au Ni5	18.3	995–1,010	7.5
Au Co5	18.2	1,010	1.8
Au Ag25 Cu5	15.2	980	8.2
Au Ag26 Ni3	15.3	1,050	8.8
Au Pt10	19.5	1,150–1,190	8.2
Au Ag25 Pt6	16.1	1,060	6.3

Arcing Contact Materials, Gold Alloys, Table 2 Application of gold and gold alloys (Doduco Data Book 1977)

Material	Applications	Typical contact forms
Fine Gold Au	Corrosion protection, improvement of fine silver contacts, contact rivets, and plug connectors	Electroplated coatings on contact rivets and small components
Electrolytic hard gold	Sliding contact tracks, rotary switches, slide switches, frequently operated plug connectors	Electroplate plated coatings on small components and printed circuits
Au Ag20	Fine contacts subject to low loads, telephone contacts for transistor circuits, frequently operated plug connectors	Solid and plated rivets, welded-on contacts
Au Ag25 Cu5Au Ag 20 Cu10	Plated contact springs and contact blades for high-grade plug connectors, relay contacts	Cladding on Cu alloys, contact rivets
Au Co5 heat treadedAu Ni5Au Ag26 Ni3	Creep-resistant fine contacts for relays, flashers, measuring instruments, electric clocks	Solid and plated rivets, welded-on contacts, claddings
Au Pt10Au Ag25 Pt5	Contacts with maximum chemical resistance for relays, measuring instruments	Contact rivets
Au Cu14 Pt 19 Ag4	Fine sliding contacts for transducers, sliding contacts for lowest voltages and currents and for maximum reliability	Bent wire components, extra-fine stamping for thin strip

(d) Alloying elements such as Ag and Cu are used to reduce cost; Pt is added to enhance the anti-welding properties.

Key Applications

Telecommunication, signal and medical connectors, switche

References

Doduco Data Book, Manual for Engineers and Businessmen, 2nd edn. (Pforzheim, 1977)

Arcing Contact Materials, Platinum-Group Metals

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

The platinum group metals are ruthenium, rhodium, palladium, osmium, iridium, and platinum. They are used as electric contacts in pure metal form, as well as alloys of the above and other elements such as silver, copper, and nickel.

Scientific Fundamentals

Contact materials in the platinum group are very expensive and used in applications requiring high reliability in chemical environments, low contact force designs, and low material transfer DC circuits. They are typically used in low-power electronic and telecommunications where signal level voltages (<5 V) and currents (<200 mA) are switched. They have better anti-sticking properties than gold alloy contacts. Ruthenium and rhodium plating are used frequently in small reed relays with low contact force of only hundredths of newton, medical connectors, and low-power signal switching. Pd-40Cu has been used in automotive flasher relays for its low DC arc transfer properties up to three million cycles.

Platinum group metals and alloys are chosen as electric contacts for several reasons:

(a) Low corrosion that keeps the contact surface clean with low surface resistance
(b) Low contact force in the milli-newtons range are sufficient to maintain reliable connectivity
(c) High melting temperature, resistant to contact welding
(d) Alloying elements to enhance properties (Tables 1 and 2)

Key Applications

Telecommunication, signal and medical connectors, switches

Arcing Contact Materials, Platinum-Group Metals, Table 1 Characteristics, processing and applications (Doduco)

Material	Characteristics	Processing	Applications
Ru	Dull gray to silvery white, very hard and brittle, resistant to all acids in the absence of oxygen, oxidizes when heated in air	Vacuum metallization, sputtering, powder metallurgy, hot forming	Powder, sheets, and wires, generally as alloying constituent. Used in reed relays
Rh	Almost silvery white, very hard and brittle, insoluble in all acids, oxidizes in air when red hot	Electrolysis, vacuum metallization, sputtering, cold-forming	Electroplated layers, alloying constituent, to a very small extent sheets and wires
Pd	Dull white, tough, and highly ductile, resistant to most non-oxidizing acids, oxidizes in air when heated to dark red	Electrolysis, vacuum metallization, dusting, cold-forming	Sheets, wires, strip, tubes, rivets
Os	Blue white, hardest platinum group metal, very brittle, resistant to non-oxidizing acids, readily oxidized in air	Powder metallurgy	Powder, as alloying constituent
Ir	Almost silvery white, very hard and brittle, resistant to all acids, oxidized by oxygen when red hot	Vapor metallization, sputtering by powder metallurgy, hot-forming	Powder mainly as alloying constituent, to a small extent as sheets
Pt	Grayish white, tough, highly ductile, resistant to all acids apart from aqua regia, HBr, resistant to oxidation even when red hot	Electrolysis, vapor metallization, sputtering, cold-forming	Sheets, wires, strip, tubes, rivets

Arcing Contact Materials, Platinum-Group Metals, Table 2 Physical properties of platinum group metals and alloys (Doduco)

Material	Density (g/cm^3)	Melting point Melting range (°C)	Electric conductivity ($m/\Omega\text{-}mm^2$)
Pt (99.9)	21.45	1773	9.4
Pt Ir5	21.5	1774–1776	4.5
Pt Ir10	21.6	1780–1785	5.6
Pt Ru10	20.6	1800	3.0
Pt Ni8	19.2	1670–1710	3.3
Pt W5	21.3	1830–1850	2.3
Pd (99.99)	12.0	1552	9.3
Pd Cu15	11.3	1370–1410	2.6
Pd Cu40	10.4	1200–1230	3.0
Pd Ni5	11.8	1455–1485	5.9

References

Doduco Data Book, *Manual for Engineers and Businessmen*, 2nd edn. (Pforzheim, Germany, 1977)

Arcing Contact Materials, Silver and Silver Alloys

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

Silver and silver alloys make up the majority of electrical contacts used in the automotive, residential, and appliance world for signal and power control. They include pure silver, fine grain silver, and silver-copper alloys that can be easily fabricated into rivet contacts for staking or welding to form electrical components in switches and relays.

Scientific Fundamentals

Silver has excellent electrical and thermal conductivity. When heated, it forms an unstable oxide that decomposes at below 200°C back to metallic form. As a result, it can maintain low surface resistance even in arc conditions that usually oxidize many other metals. It has good resistance to welding so it can be properly applied to switching in less than 20A applications when properly designed.

Silver alloys with the proper alloying components are chosen as electric contacts for several reasons (Tables 1 and 2):

Arcing Contact Materials, Silver and Silver Alloys, Table 1 Characteristics of some common Ag and Ag alloys as electrical contacts (Doduco Data Book 1977)

Material	Characteristics	Processing	Applications
Ag	Low oxidation and low surface resistance	Wire and strip forms to make rivets and buttons	Switches and relays in the 10 A range
Fine Grain Ag	The addition of .15% Ni refines the Ag grain and improves hardness without affecting oxidation	Wire and strip forms to make rivets and buttons	Switches and relays in the 15 A range with improved service life over pure Ag
AgCu	Cu lowers cost and increase hardness. This improves arc erosion but the oxidation of copper raises resistance.	Wire and strip forms to make rivets and buttons generally with less than 10 Cu	Low-cost, low-current devices and momentary switches where continuous ON currents are not required
AgPd	Pd enhances corrosion resistance, especially in a sulfur environment. This helps to maintain low contact resistance.	Wire and mini-profile tapes with only a thin layer of contact material to reduce cost	Telecommunication switches, rotary selectors, and precision potentiometers
AgMgNi	With less than.5% Ni and Mg in the alloy, this can be heat treated in air to form internal oxides that increase mechanical strength and reduce contact welding	Wire and tapes formed to contact rivets then heat treated in air to form internal precipitated oxides	Switches and relays with small size that has low contact opening forces

(a) Low oxidation rates to preserve low surface resistance.
(b) Increased hardness for low erosion wear, such as Ag-Cu with some sacrifice in contact resistance

Arcing Contact Materials, Silver and Silver Alloys, Table 2 Physical properties of silver and silver alloys (Doduco Data Book 1977)

Material	Ag	Alloying elements	Density (g/cm^3)	Melting point melting range (°C)	Electric conductivity ($m/\Omega\text{-}mm^2$)
Ag	99.95		10.5	960	60
FG Ag	99.85	.15Ni	10.5	960	60
AgCd10	90	10Cd	10.3	910	23
AgCu3	97	3 Cu	10.4	940	52
AgCu5	95	5 Cu	10.4	910	51
AgCu10	90	10 Cu	10.3	870	50
AgCu20	80	20 Cu	10.2	820	50
AgCu28	72	28 Cu	10.1	779	48
AgCuNi	75	24.5Cu 0.5Ni	10.05	1,441	43
AgMgNi	99.5	Mg,Ni	10.5	960	43
AgPd30	70	30Pd	10.8	1,220	6.4
AgPd50	50	50 Pd	11.2	1,340	3.1

(c) Improved weld resistance, such as Ag-Cd (Cd is now generally eliminated for environmental reasons).

Key Applications

Arc erosion in switches and relays

References

Doduco Data Book, Manual for Engineers and Businessmen, 2nd edn. (Pforzheim, 1977)

Arcing Contact Materials, Silver Metal Oxide

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

Silver metal oxide contacts generally refer to the common contact compositions of Ag with 10–15% metal oxides such as CdO, SnO_2, In_2O_3, or ZnO. This is a class of contact material suitable for switches, relays and contactors that must have an electrical switching life of 5,000–1,000,000 operations and still be able to maintain low contact resistance and anti-welding properties.

Scientific Fundamentals

Silver/metal oxide contacts have been in use for more than 50 years, with Ag/CdO the most commonly used in residential and industrial power controls. The switch, relay, and contactor are required to turn power on and off for many operations. Low contact resistance is achieved with the silver component, which does not oxidize to form insulating compounds in the arc heat. Anti-welding and long erosion life is achieved with the proper amount of metal oxide. Since the 1950s, Ag/CdO with 10–15% CdO has been successfully used in a few amperes (switches and automotive relays) to 2,000 A in contactors switching several hundred horsepower motors. The unique advantage of CdO in this application is its ability not only to be anti-welding, but also help to quench the arc by decomposition into Cd and O at around 900°C. This minimizes the silver melting and contact erosion. After cooling, Cd redeposits on the contact surface and forms CdO to maintain contact performance.

Ag/CdO contacts are made using powder metallurgy and internal oxidation processes.

(a) Powder metallurgy – Ag and CdO powder are mixed, pressed into green parts, and sintered at around 900°C to 85–95% density. This is then repressed to 95–98% density. The density achieved depends on the metallurgical process and is proportional to cost and performance. The same powder mix can be pressed into large billets, heated, and extruded to high-density strips and wires.

(b) Internal oxidation – AgCd alloy is made by melting Ag and Cd. This is cast and extruded into wire form or rolled to strip form. AgCd alloy, when oxidized in air or oxygen, forms CdO because the diffusion of oxygen into the metal is fast, allowing formation of CdO below the surface. The size of the CdO formed increases with distance from the surface and temperature of oxidation, and is inversely proportional to the oxygen pressure. In practice, oxidation is usually done in 100 % oxygen at atmospheric pressure or several atmospheres of oxygen pressure to accelerate the process. Variations on this include single-sided oxidation where two strips of AgCd are welded in all edges so only one side is exposed to the oxygen atmosphere. Because internal oxidation forms a gradient of CdO particle size, AgCd pellets or wires are also oxidized and then extruded into wire or strip form to homogenize the CdO distribution. Another variety is

Arcing Contact Materials, Silver Metal Oxide, Table 1 Characteristics of common Ag metal oxides

Material	Characteristics	Processing	Applications
Ag/CdO	10–20 % CdO with a balance of contact resistance, erosion life, and anti-welding properties to be used in many switches, relays, and contactors. The primary performance requirements are high number of switching cycles (5,000–1,000,000). There is no requirement for switching short circuits since that protection will be provided by the circuit breaker	Powder metal process with Ag and CdO powder using press-sinter-repress, or press-sinter-extrusion internal oxidation in air, or oxygen atmosphere at 1 to several torrs starting with a AgCd alloy. Pre-oxidation oxidizes the wire or strip and then parts are formed from this Ag/CdO form Post-oxidation is for fabrication of contact shapes with the AgCd alloy wire or strip and then internal oxidation to form the Ag/CdO mixture A silver backing is added by cladding and is required for brazing	Rivets and button shapes in residential wiring switches, appliance relays, HVAC controls Raiselays and discrete contacts brazed to copper terminals – industrial controls, starters, contactors for up to 1,000 horse power motors
Ag/SnO_2 Ag/In_2O_3 Ag/SnO_2/In_2O_3 Ag/SnO_2 with oxide additives	2–15 % total oxide	Powder metal process of press-sinter. More flexible for the addition of additive oxides to improve performance Internal oxidation of AgSnIn alloys	Rivets and button shapes in residential wiring switches, appliance relays, Automotive switches and relays, HVAC controls Raiselays and discrete contacts brazed to copper terminals – industrial controls, starters, contactors for up to 1,000 horse power motors
Ag/ZnO	8–10 % ZnO is used with good anti-welding and low resistance properties. Usually additives such as Ag_2WO_4 is required to be successful	Powder metal process of press-sinter-repress and press-sinter-extrusion	Rivets and button shapes in residential wiring switches Low current general purpose relays (15 A range)
Ag/Ni	Ag/Ni is used in switches and relays as well as circuit breakers depending on the Ni content. Ni provides some high temperature anti-welding performance, but for circuit breaker contacts, special designs on arc management is necessary	Powder metal process with Ag and Ni powder, press-sinter-extrusion followed by drawing or rolling	Residential wiring switches, and appliance relays Automotive switches and relays Low-current HVAC controls generally less than 30 A

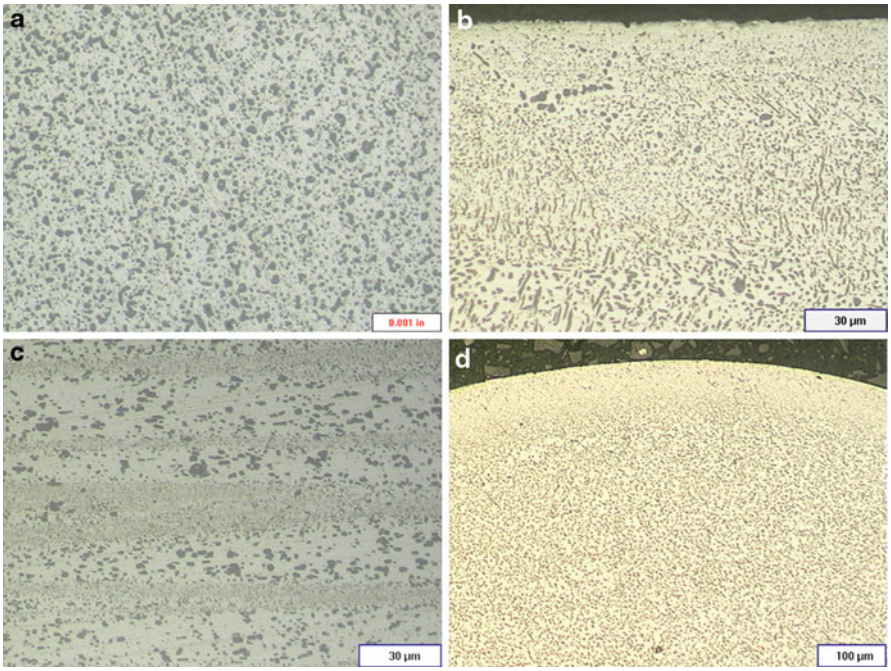
internal oxidized alloy powder. Because Ag/CdO is good in anti-welding and will not allow the contact to be brazed to an electrical terminal, a silver layer is usually cladded to the strip of Ag/CdO.

In service, the contact degradation mechanisms are: (a) surface decomposition of CdO and melting of silver to result in some splattered material loss, (b) contact welding when the Ag and CdO slowly segregate on the surface, and (c) cracking and chipping due to thermal stress from very high current arcs.

In the last 20 years, there has been increasing global concern about Cd in the electrical waste. This led to WEEE (Waste Electrical and Electronic Equipment) and ROHS (Restrictions on Hazardous Substance) regulations within the European Union and adopted by other countries. Research on and development of alternatives to CdO in electrical contacts has been ongoing by all contact manufacturers since 1970s. The most successful solution is based on Ag/SnO_2, Ag/SnO_2/In_2O_3, and Ag/ZnO. None of the three options perform as well as Ag/CdO, and more customization is necessary in contact

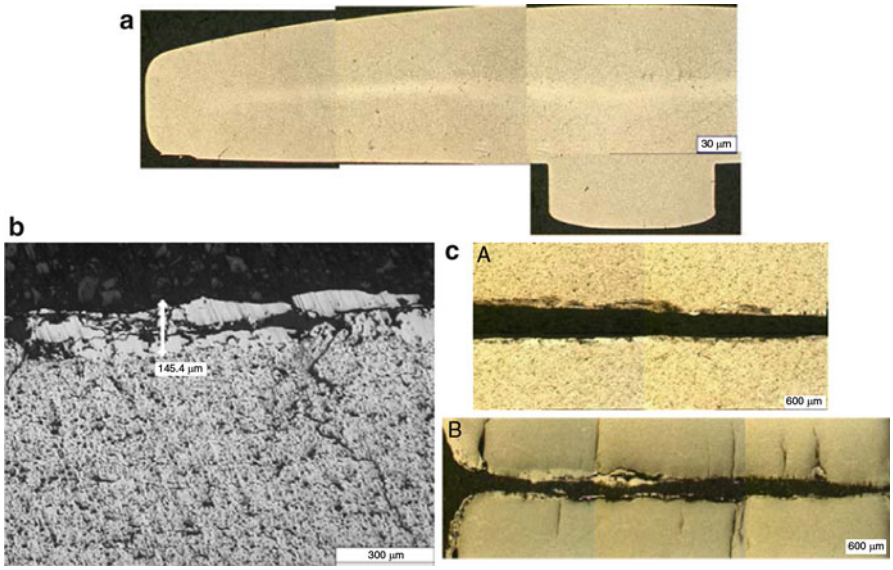
Arcing Contact Materials, Silver Metal Oxide, Table 2 Physical properties of Ag metal oxide Doduco Data Book 1977

Material	Process	Ag	Metal oxide	Typical density (g/cm^3)	Typical hardness Rockwell	Electric conductivity (m/Ω-mm^2)
AgCdO 10	PSR	90	10 CdO	10.0–10.2	F60-70	48
AgCdO 15	PSR	85	15 CdO	9.8–10.0	F65-75	43
AgCdO 20	PSR	80	20 CdO	9.7–9.9	F70-80	37
AgCdO 10	PM extrude	90	10 CdO	10.1–10.2	F78-82	50
AgCdO 15	PM extrude	85	15 CdO	10.0–10.1	F74-78	45
AgCdO 10	Int. oxidized	90	10 CdO	10.1–10.2	F71-75	50
AgCdO 15	Int. oxidized	85	15 CdO	10.0–10.1	F76-80	45
$AgSnO_2$ 10	PSR	90	10SnO_2	9.3–9.5	F50-60	38
$AgSnO_2$ 12	PSR	88	12SnO_2	9.3–9.5	F63-68	37
$AgSnO_2$ 14	PSR	86	14SnO_2	9.2–9.4	F67-75	35
$AgSnO_2$ 10	PM extrude	90	10SnO_2	9.8–10.0	F60-65	43
$AgSnO_2$ 12	PM extrude	88	12SnO_2	9.6–9.8	F65-70	41



Arcing Contact Materials, Silver Metal Oxide, Fig. 1 Ag metal contacts (**a**) 400x 85Ag15CdO made by PM press-sinter-repress; (**b**) 400x 85Ag15CdO made by single-side internal oxidation, showing gradient of CdO size; (**c**) 400x pre-oxidized grain; (**d**) 100x AgCdO 10, internal oxidized wire

material selection and device design. However, in the automotive industry, all Ag/CdO contacts have been successfully replaced by one of the above options. The cost of manufacturing is generally also higher than Ag/CdO. Contact resistance tends to be higher than equivalent Ag/CdO contacts, and minor additive oxides such as WO_3, Mo_2O_3, CuO, and Bi_2O_3 are added to improve [Leung].



Arcing Contact Materials, Silver Metal Oxide, Fig. 2 (**a**) Post-oxidized contact of 90Ag-10CdO. (**b**) 88Ag-12SnO_2 contact surface after switching 50,000 operations. Redistribution of Ag and SnO_2 plus stress cracking. Rich silver surface can cause contact welding. (**c**) Thermal stress cracks formed in endurance switching test 240VAC 90 A. The fine oxide in the IO process forms more brittle material (*A*) PSR; (*B*) Internal oxidation

Oxidation of AgSn requires the presence of other elements such as In or Bi to promote formation of internal oxides under the surface. The most successful combination is Ag/(SnO_2,In_2O_3) with the ratio of the oxides around 3:1. The high cost of indium is a concern, but the very fine oxide formed has some superior erosion properties.

Ag/ZnO also performs well to replace Ag/CdO in some lower current switching devices, such as wiring switches (Tables 1 and 2; Figs. 1 and 2).

Key Applications

Arc erosion in switches, relays, contactors

References

Doduco Data Book, Manual for Engineers and Businessmen, 2nd edn. (Pforzheim, Germany, 1977)

C. Leung, V. Behrens, A review of Ag/SnO2 contact materials and arc erosion, *24th International Conference on Electrical Contact*, (St. Malo, France, 2008)

C. Leung, E. Streicher, D. Fitzgerald, J. Cook, Contact erosion of Ag/SnO2/In2O3 made by internal oxidation and powder metallurgy, *51st IEEE Holm Conference on Electrical Contacts*, (Montreal 2001)

T. Schopf, V. Behrens, T. Honig, A. Kraus, Development of silver zinc oxide for general purpose relays, *20th International Conference on Electrical Contacts*, (Stockholm, 2000)

E. Streicher, C. Leung, D. Fitzgerald, Arc affected surface composition changes in silver tin oxide contacts, *54th IEEE Holm Conference on Electrical Contacts*, (Orlando, 2008)

Arcing Contact Materials, Silver Refractory Metals

Chi-Hung Leung
AMI Doduco, Export, PA, USA

Definition

Silver refractory metal contacts are powder metal composites of Ag/W, Ag/Mo, and Ag/WC contact materials with the silver portion offering low contact resistance and low oxidation and the refractory portion offering high melting temperature to resist arc melting and erosion. These are the most common contact materials used in residential circuit breakers as well as low-voltage industrial and commercial circuit breakers.

Scientific Fundamentals

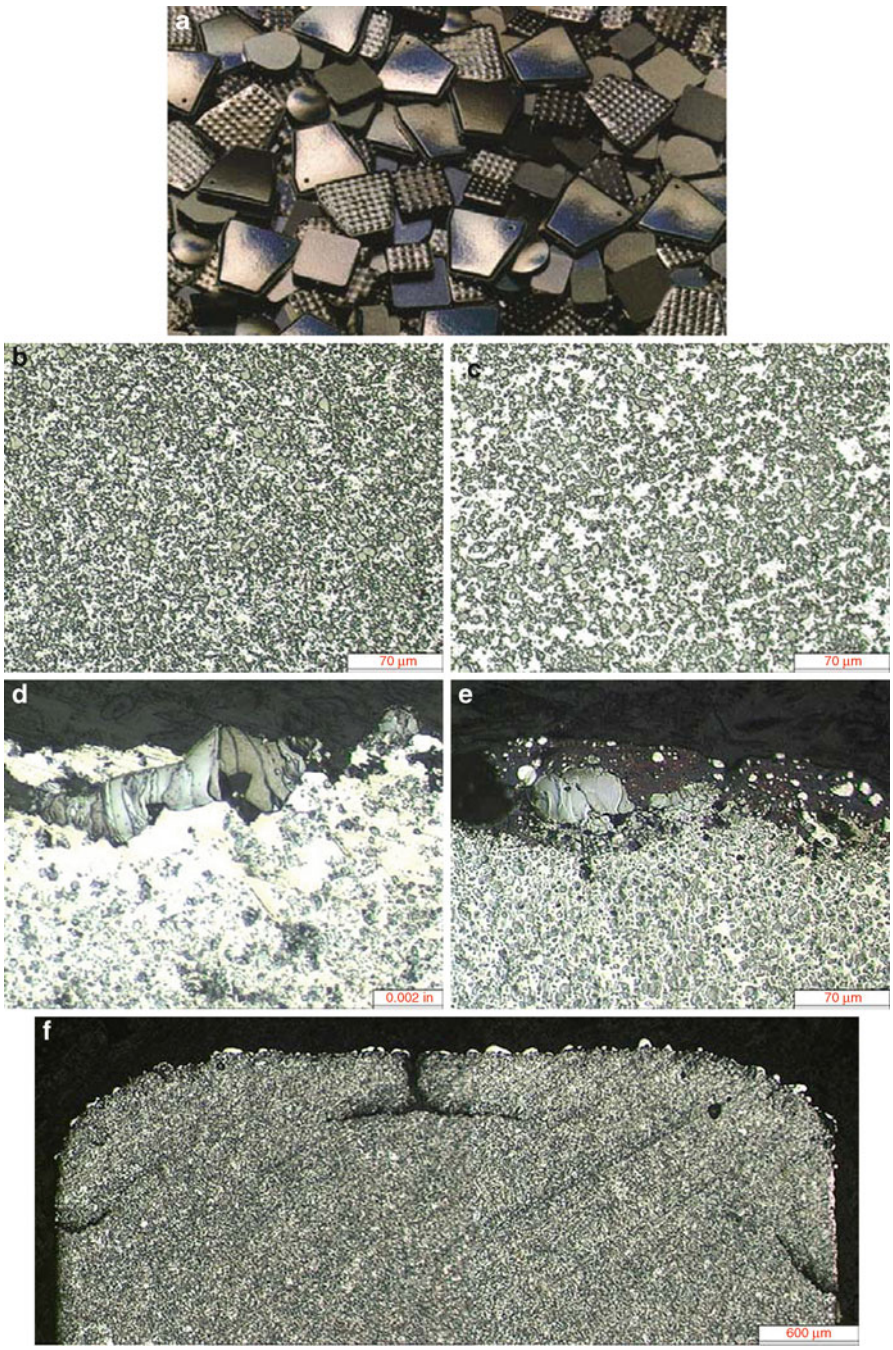
In many circuit breakers, the performance requirements of the electric contacts are (a) switching life of several thousand operations without excess erosion and causing over heating at the contact, (b) protection against overload currents of several times the rated current, and (c) a few operations at short circuit currents of several thousand amperes or tens of thousands of amperes without total destruction. The electric arc under these switching

Arcing Contact Materials, Silver Refractory Metals, Table 1 Characteristics of some common Ag refractory

Material	Characteristics	Processing	Applications
Ag/W	Common compositions are 25–50 wt% Ag, which makes it hard enough to function as arcing and main contacts in low-voltage circuit breakers. It can meet all the UL performance requirements of endurance, overload, and short circuit test when properly matched with the breaker mechanisms 80–90% Ag, balanced with W is also good contact material where over current and short circuit test requirements are not required	Powder metal process of press-sinter or press-sinter-infiltrate to near net shape with 97% density. Typical sides are.156″ diameter discs up to 1″. Surface treatments by chemical or electrochemical etching are possible to enrich the contact surface with a thin silver layer. This can improve shelf life and reduce surface resistance 80–90% Ag compositions are made by powder metal with extrusion to 99% density	UL-approved residential circuit breakers from 15 to 150 A with a single set of Ag/W moving and stationary contacts Low-voltage industrial molder case circuit breakers 600 V up to 2,000A with Ag/W contacts in combination with Ag/C contacts, or parallel main and arcing contacts Main contacts in industrial switch gears (80–90% Ag material)
Ag/Mo	40–60% Ag can have lower contact resistance than Ag/W after exposure to arc heat oxidation. However, Mo is less refractory than W and can result in higher arc erosion rates. It can also have high contact resistance in humid environments as a result of corrosion. This makes it less favored in new designs	Powder metal process of press-sinter or press-sinter-infiltrate	UL-approved residential circuit breakers Truck starters
Ag/WC	40–60% Ag is common. WC is harder than W for better arc erosion. WC, when heated by arc energy, can decompose and oxidize with less direct oxidation. However, this hard WC skeleton can make it more prone to fracture cracking. Reignition is also more severe compared with tungsten	Made by powder metal process of press-sinter-infiltrate. Carbide particles are typically in the 1–5 μm size range for optimum erosion and oxidation resistance WC is more abrasive than W and Mo, so tool wear cost is a factor in high-volume production	Molded case circuit breakers Power circuit breakers Vacuum contactors, particularly when doped with $<$2% Co to effect the chopping current, e.g., 7 kV 400 A contactors with 300,000 life operations
Ag/W/WC	35–50% Ag. The benefits of W and WC are combined to enhance certain performance requirements	Powder metal process of press-sinter-infiltrate	Residential circuit breakers of $<$100 A Molded case circuit breakers Power circuit breakers

Arcing Contact Materials, Silver Refractory Metals, Table 2 Physical properties of Ag refractory (Doduco Data Book 1977)

Material	Ag	Refractory	Typical density (g/cm^3)	Typical hardness Rockwell	Electric conductivity (m/Ω-mm^2)
AgW50	50	50	13.2	B65	36
AgW65	35	65	14.5	B85	30
AgW75	25	75	15.5	B90	26
AgMo50	50	50	10.1	B75	30
AgMo60	40	60	10.1	B85	26
AgWC40	60	40	11.8	B74	34
AgWC50	50	50	12.2	B92	29
AgWC60	40	60	12.8	B100	25



Arcing Contact Materials, Silver Refractory Metals, Fig. 1 (**a**) AgW contacts for circuit breakers; (**b**) 200x AgW65 with a mixture of fine and medium W; (**c**) 200x AgW65 with only medium tungsten; (**d**) 200x Arc-eroded contact surface with melting of W from fine powder to large solidification grains; (**e**) 200x Switch test high-resistance failure mode when contact surface is eroded to a mixture of large tungsten grain, tungsten oxide, silver beads, and silver tungsten that created high resistance; (**f**) 25x Thermal stress cracks and molten tungsten on contact surface, likely failure mode in short-circuit

conditions can create substantial oxidation and melting on the surface. It was found that a combination of good electrical conductivity, low oxidation, and high melting temperature material is necessary to meet all these requirements. Engineered powder metal composites are the only way to combine these properties into the electric contact tip.

Silver and the refractory components of W, Mo, or WC are mutually insoluble, so powder metallurgy is the method of manufacturing of the contacts. The common compositions are with 25–50 wt% Ag with the balance is the refractory component. This provides a good balance of silver to maintain low contact resistance and sufficient amount of refractory to form a strong skeleton for arc erosion.

For example, in the Ag/W system, press-sinter and press-sinter-infiltrate methods can produce parts with at least 97% dense material and less than 3 vol% porosities.

(a) Press-sinter process: Micron-size silver and tungsten powders are mixed and granulated to make a flowable powder mix. Green contacts are pressed in automated presses with uniform green density. This is then sintered in a reducing atmosphere at above 1,000°C for 30–60 min to 97% density. In this process, a finer tungsten powder as well as sintering aids such as Ni and/or Cu (total less than 1%) are necessary to get good sintering. Dimension shrinkage is around 10%.
(b) Press-sinter-infiltrate: Green contacts are sintered at 1,100–1,350°C for less than 60 min to a 1–2% shrinkage. This is then infiltrated with molten silver to 98% density, followed by sintering and/or infiltration to 98% density with less than 3 vol% porosities.

Particle size, uniformity of tungsten distribution, process temperatures are ways to modify electrical performance of metallurgical designs. Electric contacts of Ag/Mo, Ag/WC, and Ag/(W + WC) are similar to Ag/W. Molybdenum has a lower melting temperature than W and also tends to have higher arc erosion rates. However, an arc erosion mixture of molybdenum oxide and silver has lower contact resistance than Ag/W, making it easier to keep the circuit breaker cool. The disadvantage of molybdenum is that it is more prone to atmospheric corrosion by humidity. This can, in turn, increase contact resistance at the contact surface and cause tripping of the circuit breaker. WC tends to increase hardness and resistance to erosion, however, it has higher tendency to reignition, making arc extinction more difficult.

In service, the contact degradation mechanisms are: (a) high surface resistance as a result of tungsten oxide and silver tungstate formation on the surface in each switching operation, (b) excess erosion especially in overload and short circuit current operations, and (c) cracking and chipping as a result of severe thermal stress under very high current for short durations of 10–50 ms (Tables 1 and 2; Fig. 1).

Key Applications

Arc erosion in circuit breaker contacts

References

Doduco Data Book, Manual for Engineers and Businessmen, 2nd edn. (Pforzheim, 1977)

Arc-PVD

► Cathodic Arc Technology

Areal Filter

► Filtration of Surface Measurement Data

Articular Cartilage as a Bearing Material, An Engineering Perspective

John Fisher
Institute of Medical and Biological Engineering, School of Mechanical Engineering, University of Leeds, Leeds, UK

Definition

Articular cartilage and the underlying bone form the bearing surface and primary load carrying structures in the natural joint. The muscles control the forces applied to the joint, the ligaments and intrinsic joint surface geometries provide the constraint and stability, while the synovial fluid and intrinsic tissue fluid provides the lubrication. These materials and structures form a complex bioengineering tribological system, which enables bodies to move efficiently with low friction and wear. As we age the tissue structures and engineering system gradually degrade and need to be enhanced, substituted, or replaced.

Scientific Fundamentals

Structure and Components of the Synovial Joint – An Engineering Perspective

The bones and articular cartilage attached to the ends of the bones are the main structures responsible for transmitting loads through the body and across the joints (refer to ► Tribological Design of Natural Joints, an Anatomical Perspective). The forces across the joint and acting through the bone and cartilage structures are between two and four times body weight in the knee, four to six times body weight in the hip, and even higher during extreme activities such as jumping. The forces acting across the cartilage and bone in the joint are a combination of the direct weight of the body segments, the muscle forces associated with movements and stability acting across the joint, and the dynamic forces associated with time-dependent movements and inertia. For example, the force across the hip joint during a static one-leg stance is approximately two times body weight, during mid-stance phase of walking three times body weight, and at heal strike four times body weight. The muscles forces across the joint control the motion, but also act to stabilize the joint and keep the bones and cartilage in the correct relative position. The ligament structures in the joint also act to stabilize the joint, keeping the bearing surfaces, the cartilage, and bone in the correct position. In the knee, the meniscus acts partly as a ligament to stabilize the joint, but also acts as a load-bearing spacer to make the joint more conforming and reduce the contact stresses in the bone and cartilage. The joints have different geometries, a conforming ball and socket in the hip, an ellipsoid geometry in the shoulder, and two separate ellipsoid condylar geometries in the knee. These geometries determine the types of motions (rotations and translations) possible in each of the joints. All these structures and components contribute to the effective working of the synovial joint as a bioengineering tribological system. Failure of any one component can result in adverse loading or contact stresses and damage and degradation to the articular cartilage and bone. For example, damage or deterioration of the meniscus in the knee increases the contact stresses in the underlying cartilage and bone. Rupture or damage to a ligament can produce instability, malalignment, adverse kinematics, and increase the loads and contact stresses in the articular cartilage and bone. Elevation of the contact stresses can result in direct damage, fatigue, elevated wear of the articular cartilage, or disruption of biological function.

Cartilage and Bone as a Composite Structure – The Contact Mechanics of the Bearing Surfaces

The bearing surface of a natural joint comprises cartilage and bone as a composite structure. The articular cartilage is a thin compliant layer that covers the articulating surfaces of the much stronger and stiffer bone. The thin layer of articular cartilage (1–3 mm thick) with a low aggregate modulus of approximately 1 MPa is structurally integrated with the much stiffer underlying subchondral cortical bone, with an elastic modulus of approximately 10 GPa. This creates a single composite structure. It is the composite structure that is able to withstand and transmit the contact stresses across the joint, typically in the range 1–5 MPa. This is essentially a thin compliant layer contact. The much stiffer and stronger underlying bone provides the structural support for the softer and more compliant cartilage layer. The thin layer construct transmits loads through a hydrostatic stress state, with reduced shear stresses in the solid phase of the cartilage. Without the integrated support of the underlying subchondral bone in the thin layer contact, the articular cartilage would not be able to transmit or withstand the contact stresses applied in the natural joint (Jin et al. 1991). When considering the tribology and lubrication it is necessary to consider the biphasic and porous nature of the articular cartilage (Jin et al. 1992) in the thin layer contact. The biphasic nature of articular cartilage ensures load carriage by the fluid phase, thus generating internal hydrostatic fluid pressure or biphasic lubrication, which reduces friction and wear.

Friction and Lubrication of Cartilage

The importance of the fluid phase of the articular cartilage in controlling load carriage, lubrication, and friction was first recognized by McCutchen 1960, who described it as both self-pressurized hydrostatic lubrication and weeping lubrication. This was defined and developed by Mow and colleagues in the form of biphasic theory (Mow et al. 1980, 1984) and, more recently, the term *biphasic lubrication* has been used, which captures the importance of fluid pressurization and load carriage by the fluid phase (Forster et al. 1996; Krishnan et al. 2004). The frictional force in articular cartilage, which is primarily determined by the biphasic lubrication and proportion of load carried by pressurization of the fluid phase, is dependent on the time for which the load (or stress) is applied and the level of load or contact stress (Forster et al. 1996, 1999; Katta et al. 2007). The friction coefficient can range from 0.01 and 0.3 depending on the proportion of load carriage by the fluid phase. The friction coefficient is dependent

on the glycosaminoglycan (GAG) content and permeability of the cartilage and its ability to retain the fluid pressurization (Katta et al. 2008). (Refer to ► Biphasic Lubrication).

Other lubrication mechanisms also play a role in controlling friction (Murakami et al. 1998). Proteins (Radin et al. 1970; Swann et al. 1981; Bell et al. 2006) and lipids (Hills 2003) all have a role to play in boundary layer lubrication particularly at low stress level. These combine as a biphasic surface amorphous layer to enhance biphasic lubrication (Graindorge et al. 2005). Surface fluid film lubrication mechanisms also have a role to play in reducing friction, including squeeze film lubrication and elasto-hydrodynamic fluid lubrication (refer to ► Elastohydrodynamic Lubrication of Natural Synovial Joints) (Dowson et al. 1969). (► Introduction to Biotribology).

Wear and Degradation of Cartilage and Bone

The sophisticated lubrication regimes protect articular cartilage and underlying bone. However, both cartilage and bone degrade with age and lose their load-carrying and lubrication properties, which leads to wear. Additionally, adverse conditions that elevate loads and stresses cause accelerated fatigue damage and wear. There is, however, much less knowledge of wear and degradation of cartilage (Lipshitz et al. 1995; Forster et al. 1999). McCann et al. have defined the dependency of cartilage wear on contact stress and frictional shear stress (McCann et al. 2009) and demonstrated the importance of the meniscus in reducing articular wear and contact stress in the knee (McCann et al. 2009).

Approaches to Enhancement and Substitution

Approaches to enhancing lubrication, improving tribology, and reducing friction and wear include injectable biological lubricants (Bell et al. 2006; Forsey et al. 2006) and tissue substitutions through biomaterials or biological scaffolds (Oka et al. 2000; Northwood et al. 2007; McCann et al. 2009; Plainfosse et al. 2007). As use of tissue substitution and tissue engineering solutions increased, the tribology of the repaired whole natural joint system will be critical to successful clinical applications.

Cross-References

► Introduction to Biotribology
► Tribological Design of Natural Joints, an Anatomical Perspective

References

C.J. Bell, E. Ingham, J. Fisher, Influence of hyaluronic acid on the time-dependent friction response of articular cartilage under different conditions. Proc. Inst. Mech. Eng. J Eng. Med. **220H**, 23–31 (2006)

D. Dowson, V. Wright, M.D. Longfield, Human joint lubrication. Biomed. Eng. **4**, 160–165 (1969)

R.W. Forsey, J. Fisher, J. Thompson, M.H. Stone, C. Bell, E. Ingham, The effect of hyaluronic acid and phospholipid based lubricants on friction within a human cartilage damage model. Biomaterials **27**, 4581–4590 (2006)

H. Forster, J. Fisher, The influence of loading time and lubricant on the friction of articular cartilage. Proc. Inst. Mech. Eng. J. Eng. Med. **210H**, 109–119 (1996)

H. Forster, J. Fisher, The influence of continuous sliding and subsequent surface wear on the friction of articular cartilage. Proc. Inst. Mech. Eng. J. Eng. Med. **213H**, 329–345 (1999)

S. Graindorge, W. Ferrandez, Z.M. Jin, E. Ingham, C. Grant, P. Twigg, J. Fisher, Biphasic surface amorphous layer lubrication of articular cartilage. Med. Eng. Phys. **27**, 836–844 (2005)

B.A. Hills, R.W. Crawford, Normal and prosthetic synovial joints are lubricated by surface-active phospholipid: a hypothesis. J. Arthroplasty **18**, 499–505 (2003). 85

Z.M. Jin, D. Dowson, J. Fisher, Stress analysis of cushion form bearings for total hip replacements. Proc. Inst. Mech. Eng. J. Eng. Med. **205H**, 219–226 (1991)

Z.M. Jin, D. Dowson, J. Fisher, The effect of porosity of articular cartilage on the lubrication of a normal hip joint. Proc. Inst. Mech. Eng. J. Eng. Med. **206H**, 117–124 (1992)

J. Katta, S. Pawaskar, Z. Jin, E. Ingham, J. Fisher, Effect of load variation on the friction properties of articular cartilage. Proc. Inst. Mech. Eng. J. Eng. Tribol. **221J**, 175–181 (2007)

J. Katta, T. Stapleton, E. Ingham, Z. Jin, J. Fisher, The effect of glycosaminoglycan depletion on the friction and deformation of articular cartilage. Proc. Inst. Mech. Eng. J. Eng. Med. **222H**, 1–11 (2008)

R. Krishnan, M. Kopacz, G.A. Ateshian, Experimental verification of the role of interstitial fluid pressurization in cartilage lubrication. J. Orthop. Res. **22**, 565–570 (2004)

H. Lipshitz, R. Etheredge 3, M.J. Glimcher, In vitro wear of articular cartilage. J. Bone Joint Surg. **57A**, 527–534 (1975)

L. McCann, E. Ingham, Z. Jin, J. Fisher, An investigation of the effect of conformity of knee hemiarthroplasty designs on contact stress, friction and degeneration of articular cartilage: a tribological study. J. Biomech. **42**, 1326–1331 (2009a)

L. McCann, E. Ingham, Z. Jin, J. Fisher, Influence of the meniscus on friction and degradation of cartilage in the natural knee joint. Osteoarthr. Cartil. **17–8**, 995–1000 (2009b)

C.W. McCutchen, Sponge-hydrostatic and weeping bearings. Nature **184**, 1284–1285 (1959)

V.C. Mow, S.C. Kuei, W.M. Lai, C.G. Armstrong, Biphasic creep and stress relaxation of articular cartilage in compression: theory and experiments. J. Biomech. Eng. **102**, 73–84 (1980)

V.C. Mow, M.H. Holmes, W.M. Lai, Fluid transport and mechanical properties of articular cartilage: a review. J. Biomech. **17**, 377–394 (1984)

T. Murakami, H. Higaki, Y. Sawae, N. Ohtsuki, S. Moriyama, Y. Nakanishi, Adaptive multimode lubrication in natural synovial joints and artificial joints. Proc. Inst. Mech. Eng. J. Eng. Med. **212H**, 23–35 (1998)

E. Northwood, J. Fisher, R. Kowalski, Investigation of the friction and surface degradation of innovative chondroplasty materials against articular cartilage. Proc. Inst. Mech. Eng., J. Eng. Med. **221H**, 263–279 (2007)

M. Oka, K. Ushio, P. Kumar, K. Ikeuchi, S.H. Hyon, T. Nakamura, Development of artificial articular cartilage. Proc. Inst. Mech. Eng. J. Eng. Med. **214H**, 59–68 (2000)
M. Plainfosse, P.V. Hatton, A. Crawford, Z.M. Jin, J. Fisher, Influence of the extracellular matrix on the frictional properties of tissue-engineered cartilage. Biochem. Soc. Trans. **35**, 677–679 (2007)
E.L. Radin, D.A. Swann, P.A. Weisser, Separation of a hyaluronate-free lubricating fraction from synovial fluid. Nature **228**, 377–378 (1970)
D.A. Swann, H.S. Slayter, F.H. Silver, The molecular structure of lubricating glycoprotein-I, the boundary lubricant for articular cartilage. J. Biol. Chem. **256**, 5921–5925 (1981)

Articulating Joint

► Coatings for Biomedical Applications

Artificial Knee Contact Mechanics

► Contact Conditions in the Artificial Knee

Artificial Knee Joint

► Lubrication in Knee Prostheses

Artificial Synovial Fluid

Jian-Hua Zhang
School of Mechatronics and Automation, Shanghai University, Shanghai, People's Republic of China

Synonyms

Biomimetic synovial fluid; Bionic synovial fluid

Definition

Artificial synovial fluid, a substitute for synovial fluid, is a kind of biomimetic lubricant for human joint therapy or artificial joint lubrication.

Scientific Fundamentals

Mechanism of Synovial Fluid

Synovial fluid is a viscous, non-Newtonian fluid found in the cavities of synovial joints. With its yolk-like consistency (*synovial* partially derives from *ovum*, Latin for *egg*), the principal role of synovial fluid is to reduce friction between the articular cartilage of synovial joints during movement (Balazs 1974). Its functions are reducing friction by lubricating the joint, absorbing shocks, and supplying oxygen and nutrients to and removing carbon dioxide and metabolic wastes from the chondrocytes within articular cartilage.

The functions of the synovial fluid include:

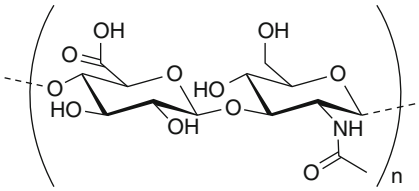
- *Reduction of friction* – Synovial fluid lubricates the articulating joints.
- *Shock absorption* – As a dilatant fluid, synovial fluid is characterized by the rare quality of becoming more viscous under applied pressure. The synovial fluid in diarthrotic joints becomes thick the moment shear is applied in order to protect the joint and subsequently thins to normal viscosity instantaneously to resume its lubricating function between shocks.
- *Nutrient/waste transportation* – The fluid supplies oxygen and nutrients and removes carbon dioxide and metabolic wastes from the chondrocytes within the surrounding cartilage.

Synovial fluid can degrade due to inflammation. Degraded synovial fluid can be classified into several different stages, such as normal, noninflammatory, inflammatory, and septic.

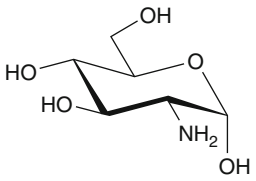
Artificial Synovial Fluid – Hyaluronan

Hyaluronan $(C_{14}H_{21}NO_{11})_n$, also called hyaluronic acid or hyaluronate, is a high-molecular-mass polysaccharide found in the extracellular matrix, especially of soft connective tissues. It is synthesized in the plasma membrane of fibroblasts and other cells by addition of sugars to the reducing end of the polymer, whereas the nonreducing end protrudes into the pericellular space. The polysaccharide is catabolized locally or carried by lymph to lymph nodes or the general circulation, from where it is cleared by the endothelial cells of the liver sinusoids. The overall turnover rate is rapid for a connective tissue matrix component (Laurent and Fraser 1992).

Hyaluronan is a negatively charged linear polysaccharide that consists of repeated linear disaccharide units of β-D-glucuronic acid and *N*-acetyl-β-D-glucuronic residues linked at 1,4 and 1,3 positions (Fig. 1). The segmental mobility of the Hyal chain is more restricted than that of similar polysaccharides owing to local hydrogen bonds. This structure helps to maintain and attract water molecules, form a collagenous mesh, and bind proteoglycans for cartilage.



Artificial Synovial Fluid, Fig. 1 Hyaluronan



Artificial Synovial Fluid, Fig. 2 Glucosamine

As one artificial synovial fluid for joint injection therapy, hyaluronan has also been recognized in clinical medicine. A concentrated solution of hyaluronan (10 mg/ml) has, through its tissue protective and rheological properties, become a device in ophthalmic surgery. It is unique among glycosaminoglycans in that it is nonsulfated, forms in the plasma membrane instead of the Golgi apparatus, and can be very large, with its molecular weight often reaching the millions.

Artificial Synovial Fluid – Glucosamine

As artificial synovial fluid, glucosamine ($C_6H_{13}NO_5$) is injected for cartilage repair, directly improving the tribological performance of human joints (Fig. 2). Glucosamine is an amino sugar and a prominent precursor in the biochemical synthesis of glycosylated proteins and lipids. It is part of the structure of the polysaccharides chitosan and chitin, which compose the exoskeletons of crustaceans and other arthropods and cell walls in fungi and many higher organisms. Glucosamine is utilized in the production of glycosaminoglycans (GAGs), which then combine with hyaluronic acid (HA) to form large proteoglycans in the matrix of cartilage. As such, glucosamine is a fundamental molecule for the synthesis and structure of cartilage. Because glucosamine is the major structural polysaccharide of cartilage in joints, a reduced amount of glucosamine in cartilage is a feature of osteoarthritis.

Glucosamine, which is needed by the body to make glycosaminoglycans, is used to treat osteoarthritis of the knee, and can be administered by mouth, intramuscular injection, or by direct injection into the knee. Clinical studies show that glucosamine and chondroitin (GCS) can help patients to generate a higher synovial fluid viscosity than those on placebo. It was also found that patients on placebo had a decreased viscosity while patients taking GCS had not changed (Clegg et al. 2006).

Biomimetic Synovial Fluid

Biomimetic synovial fluid is a kind of artificial synovial fluid developed for the lubrication of artificial joint and the treatment of particle-induced osteolysis (Fig. 3).

Artificial Synovial Fluid, Fig. 3 Biomimetic synovial fluid

The human joint is regarded as a biological system and a healthy human joint can last for more 70 years with minimum friction and wear. Synovial fluid in native joints functions as a biomechanical lubricant, lowering the friction and wear of articulating cartilage in synovial joints. It is found that synovial fluid, an aqueous electrolyte solution rich in protein, lipids, and hyaluronan, in combination with articular cartilage, provides low-friction coefficient towards human joints. However, following arthroplasty, such biological system is destroyed, the synovial membrane is reformed and the artificial materials are poorly lubricated by pseudo-synovial fluid. In laboratory tests with prostheses, it has also been demonstrated that the lubricant has a profound effect on wear (Ahlroos and Saikko 1997) and lubrication regime (Jalali-Vahid et al. 2001). This is underscored when comparing no lubricant, water, and bovine serum lubrication conditions. Therefore, the lubricant is regarded as one of the key factors towards tribological performance of artificial joints.

As an ideal substitute, biomimetic synovial fluid has been developed for artificial joint lubrication and treatment of particle-induced osteolysis. Thus, the biomimetic synovial fluid contains both lubrication and therapeutic additives.

For example, hyaluronan is added as a component to adjust the viscosity, viscoelasticity, and flow properties of the lubricants, thus reducing the friction. Albumin and γ-globulin, which are the abundant proteins in synovial fluid, are added for joint nutrition as well as generation of protein film lubrication to prevent the material transfer and friction reduction. Another component, alendronic acid sodium (in medications called bisphosphonates (BPs)), which to hydroxyapatite in bone and acts as an inhibitor of bone resorption. BPs are usually used for the treatment of osteitis deformans, postmenopausal osteoporosis, and hypercalcemia related to malignancy. Therefore, alendronic acid sodium can be chosen as a therapeutic component for particle-induced osteolysis (Hua et al. 2007, 2010).

As the example shows above, biomimetic synovial fluid can be regarded as a synthetic synovial fluid, containing lubricant, nutrient, and therapeutic elements. Several characteristics need verification before medical application:

Biocompatibility tests are related to the behavior of biomimetic synovial fluid in various contexts and should be verified in accordance with ISO 10993 or other similar standards. These tests do not determine the biocompatibility of a material, but they do constitute an important step toward animal testing and finally clinical trials that will determine the biocompatibility of the material in a given application, and thus medical devices such as implants or drug-delivery devices (ISO 10993–1 2003).

Tribological tests are related to the lubrication performance of the biomimetic synovial fluid. General tribological tests, such as pin-on-disk tests (ASTM G99-05 2010) and four ball wear tests (ASTM D4172-94 2010), in accordance with ASTM or other similar standards, are performed to determine the wear protection properties of the lubricant. Furthermore, for artificial joint lubrication applications, due to the multidirectional motion and load factors of human joints, biomimetic synovial fluids should be tested according to ISO prostheses wear test methodology, such as artificial hip joint (ISO 14242 2012), knee joint (ISO 14243 2009), and spine (ISO 18192–2 2010).

Tribo-electrochemical tests, or tribo-corrosion tests, are experiments that describe the synergy between tribological and electrochemical processes of material in an electrolyte solution. This test method was applied by Yan et al. to investigate the tribo-electrochemical characteristics of CoCrMo (Yan et al. 2006). As an organic fluid, biomimetic synovial fluid should undergo tribocorrosive evaluation, especially if used for artificial joints.

Key Application

Artificial Synovial Fluid for Joint Injection Therapy

Currently, artificial synovial fluids, such as hyaluronan, glucosamine, and biomimetic synovial fluids are mostly applied in the joint injection therapy process. A joint injection (intra-articular injection) is a procedure used in the treatment of inflammatory joint conditions, such as rheumatoid arthritis, psoriatic arthritis, gout, tendinitis, bursitis, and osteoarthritis (Lavelle et al. 2007). Figure 4 shows an example of knee joint injection therapy with artificial synovial fluids. It is a treatment approved by the Food and Drug Administration (FDA) to treat pain associated with osteoarthritis of the joints in patients who have failed to see results from palliative care or physical therapy and exercise. Artificial synovial fluid involved in joint injection therapy is developed based on the natural synovial fluids, to replace bursa fluids and deliver anti-inflammatory agents into joints.

Intra-articular injections of hyaluronan are known as *viscosupplementation*. Balazs (1968) has advocated the use of such artificial synovial fluid since 1960s. The goal of viscosupplementation is to recreate the environment of a healthy joint in an effort to encourage the joint to escape the negative feedback cycle of osteoarthritis progression. Once normal viscoelastic properties are returned to the joint fluid, the cartilage lining the joint will increase production of healthy synovial fluid.

Artificial Synovial Fluid, Fig. 4 Knee joint injection therapy with artificial synovial fluids

By viscosupplementation, this fluid helps to reduce friction from roughened cartilage and cushions the joint, producing an analgesic effect. It has also been suggested that hyaluronan has positive biochemical effects on cartilage cells. However, some placebo-controlled studies have cast doubt on the efficacy of hyaluronan injections, and hyaluronan is recommended primarily as a last alternative before surgery (Karlsson et al. 2002).

Cross-References

► Lubricant Viscosity
► Lubrication with a Non-Newtonian Fluid

References

T. Ahlroos, V. Saikko, Wear of prosthetic joint materials in various lubricants. Wear **211**, 113–119 (1997)

ASTM D4172-94, *Standard Test Method for Wear Preventive Characteristics of Lubricating Fluid (Four-Ball Method)*. Annual Book of ASTM Standards, ASTM International, West Conshohocken, PA, US (2010)

ASTM G99-05, *Standard Test Method for Wear Testing with a Pin-on-Disk Apparatus*. Annual Book of ASTM Standards, ASTM International, West Conshohocken, PA, US (2010)

E.A. Balazs, Viscoelastic properties of hyaluronic acid and biological lubrication. Univ. Mich. Med. Cent. J. **9**, 255–259 (1968)

E.A. Balazs, The physical properties of synovial fluid and the special role of hyaluronic acid, in *Disorders of the Knee*, ed. by A. Helfet (Lippincott, Philadelphia, 1974), pp. 63–75

D.O. Clegg, D.J. Reda, C.L. Harris, D. Pharm, M.A. Klein, J.R. O'Dell, Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. N. Engl. J. Med. **354**, 795–808 (2006)

Z.K. Hua, S.H. Su, J.H. Zhang, Tribological study on new therapeutic bionic lubricants. Tribol. Lett. **28**, 51–58 (2007)

Z.K. Hua, P. Gu, J.H. Zhang, Tribological and electrochemical studies on biomimetic synovial fluids. Sci. China Technol. Sci. **53**, 2996–3001 (2010)

ISO 10993–1, *Biological Evaluation of Medical Devices* (ISO, Geneva, 2003)

ISO 14242, *Implants for Surgery – Wear of Total Hip-Joint Prostheses* (ISO, Geneva, 2012)

ISO 14243, *Implants for Surgery – Wear of Total Knee-Joint Prostheses* (ISO, Geneva, 2009)

ISO 18192–2, *Implants for Surgery – Wear of Total Intervertebral Spinal Disc Prostheses* (ISO, Geneva, 2010)

D. Jalali-Vahid, M. Jagatia, Z.M. Jin, D. Dowson, Prediction of lubricating film thickness in UHMWPE hip joint replacements. J. Biomech. **34**, 261–266 (2001)

J. Karlsson, L.S. Sjögren, L.S. Lohmander, Comparison of two hyaluronan drugs and placebo in patients with knee osteoarthritis. A controlled, randomized, double-blind, parallel-design multicentre study. Rheumatology **41**, 1240–1248 (2002)

T.C. Laurent, J.R. Fraser, Hyaluronan. Fed. Am. Soc. Exp. Biol. J. **6**, 2397–2404 (1992)

W. Lavelle, E.D. Lavelle, L. Lavelle, Intra-articular injections. Anesthesiol. Clin. **25**, 853–862 (2007)

Y. Yan, A. Neville, D. Dowson, Biotribocorrosion–an appraisal of the time dependence of wear and corrosion interactions: I. The role of corrosion. J. Phys. D: Appl. Phys. **39**, 3200–3205 (2006)

Ashless Dispersants

► Dispersant Additives

Ashless Phosphate Esters

W. David Phillips
Manchester, UK

Synonyms

Acid phosphates; Antiwear additives; Extreme pressure additives; Fire-resistant hydraulic fluids; Trialkyl phosphates; Triaryl phosphates

Definitions

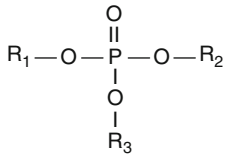
The term "phosphate ester" covers a huge number of compounds with a wide range of properties but all contain the core structure of a phosphorus atom surrounded by four oxygen atoms, as shown in Fig. 1 below:

This structure obviously precludes a discussion of the metal thiophosphates (e.g., zinc dialkyl dithiophosphate, ZDDP). However, of similar structure to the phosphate esters are the thionophosphates (or phosphorothionates) where the –P = O bond is replaced by –P = S and mention will be made of these compounds.

– *Antiwear additives (AW)* are used to provide reduced wear and friction under low–medium loads (or what may be called mixed-friction conditions) as are found in the operation of hydraulic systems, turbine and compressor lubrication etc. They are used in mineral oil and most synthetic lubricant types.

  The compositions of phosphates used as this type of additive are usually neutral alkyl or aryl esters e.g., where $R_{1–3}$ are all either $C_{4–10}$ alkyl or $C_{6–14}$ aryl/substituted aryl (Phillips 2009a; Barcroft and Daniel 1964; Forbes and Silver 1970).

– *Extreme pressure additives (EP)* are used to reduce wear and to increase the load at which seizure (or scuffing) occurs under very high applied loads (usually called boundary lubrication conditions). These conditions are found, for example, in gears and metal working operations (e.g., cutting, drilling and tapping). While the neutral esters do not, by themselves, provide good performance under these more severe conditions, the alkyl esters can behave synergistically with



Ashless Phosphate Esters, Fig. 1 The basic structure of a phosphate ester, where R_1, R_2 and R_3 may be the same or different and are usually either hydrogen, alkyl or aryl groups as indicated below

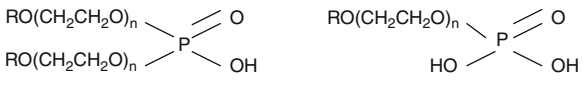
sulphur-containing additives to provide the necessary high load-carrying performance (Phillips 2009a).

The replacement of the –P = O function by –P = S, as in the thionophosphates, results in a significant improvement in extreme pressure performance. Of these compounds the most widely used is triphenylphosphorothionate (TPPT) which is found in some grease and circulatory oil formulations.

Alternatively, for oil-based products in circulatory applications (hydraulic oils turbine oils etc.), long chain amine phosphates may be used. Such products would have a general formula $(R_1O)_{1-2}$ PO· OH $(H_2NR_2)_{2-1}$ where R_1 is typically C_{4-8} and R_2, alkyl C_{11-14}.

For water-based applications where high load carrying is required, particularly in metal working, the above products are not suitable because of their poor solubility. As a result the phosphate chemistry used in soluble oils, synthetic and semi-synthetic metal working formulations, is different. Here polyoxyethyleneoxy- acid phosphate esters are used – normally as the amine salts. The presence of base assists in providing corrosion inhibition and in keeping the pH of the fluid alkaline, thereby preventing microbial growth.

The products are manufactured from alcohols which have been reacted with ethylene oxide, and dependent on the number of units of ethylene or propylene oxide present and the type of amine used for neutralisation, the product may be oil soluble, water soluble or dispersible in both media. Products used in this application would typically be mixtures of mono and di-acid phosphates of structure:



where "R" is an alkyl (C_{8-10}), alkylaryl (C_{18}), dialkylaryl (C_{24}) or aryl (C_6) and "n" is the number of moles of ethylene oxide used (Placek and Shankwalkar 1994).

– *Fire-resistant hydraulic fluids* a hydraulic fluid transfers energy (in the form of pressure) from one part of a system to another through the fluid itself. Fluids which do not compress significantly under pressure are the most efficient and can be considered for use. Mineral oil is, of course, widely used. However some applications involve high temperatures and if the hydraulic fluid was to escape from the system it could ignite with potentially dangerous results. For such applications "fire-resistant" fluids are used. These fluids are more difficult to ignite but if they do, they should preferably not propagate flame. Phosphate esters are one class of fluid that shows significant resistance to ignition and self-extinguishes rapidly when it is no longer in contact with the ignition source (Phillips 2011).

The phosphates which are used for this application are either neutral C_4 alkyl phosphates (for commercial aviation) or neutral C_{6-14} substituted aryl phosphates in general industrial applications. The alkyl phosphates are used in aircraft because they possess much better low temperature properties and remain liquid when exposed to the extreme cold at high altitude. The aryl phosphates would solidify under these conditions.

Scientific Fundamentals

Mechanism of Phosphate Esters as Antiwear and Extreme Pressure Additives

The mechanism of phosphate esters as AW/EP additives has long been debated as it is complex and influenced by many variables. A summary of the most widely accepted theories follows:

In the case of neutral phosphates, the phosphate molecule adsorbs onto the metal via the –P = O bond or, more likely, via any existing acidic –P–OH bonds. (There is always a small amount of residual acid present.) Once attached to the surface, hydrolysis of other–P–OR groups takes place (probably arising from water at the metal surface) to produce –P–OH bonds which then react with the metal surface to form iron salts. Hydrolysis of the neutral soap in solution will result in the availability of additional –P–OH groups to maintain the reaction (Barcroft and Daniel 1964; Forbes and Silver 1970).

Acid phosphates already have many –P–OH bonds present and therefore react more readily with the surface. While the presence of a long hydrocarbon chain on the acid phosphate will undoubtedly assist in forming a thicker film, the presence of two –P–OH bonds on the same phosphorus atom may also facilitate the formation of a polyphosphate layer. Indeed, the formation of polyphosphate has been the focus of more recent investigations (Placek and Shankwalkar 1994; Saba and

Forster 2002). Saba and Forster (2002) suggests that at high temperatures –C–O– bonds are cleaved while the phosphate is attached to the surface eliminating aryl radicals (which can then form amorphous carbon). A "lattice of cross-linked $–PO_3$ is then formed with the metal surface."

The function of the carbon may be to act as the actual lubricant which is kept on the surface by the polyphosphate acting as a "binder." Wear of the film is not regarded as a problem as it appears to be self-healing due to diffusion of Fe ions through the polyphosphate layer where reaction with the phosphate continues (Saba and Forster 2002).

The presence of carbon in the reaction layer could account for some reports of the presence of iron phosphide on the surface; this material being produced by the reduction of iron phosphate by carbon at high temperatures.

The formation of amine salts (of, for example, the alkyl and aryl polyoxyethleneoxy- acid phosphates) results in an increase in activity, possibility because of improved stability of the ion and improved adsorption on the metal surface.

While the above mechanisms may help clarify what is happening at the metal surface they do not explain why the formation of metal soaps or salts – or even carbon - results in reduced wear and increased load-carrying. Generally speaking the process is a result of either physical separation of the rubbing surfaces (to reduce friction) or modification of one (or both) surfaces to produce a lower friction coating - although this may initially involve some wear. The process is, therefore, temperature dependent (which, in turn varies with the load and relative speeds of the contacting surfaces). Thus, at very low loads, the physical adsorption of a phosphate film may be sufficient to keep the surfaces apart but as the load and temperatures increase, the film starts to react chemically with the surface initially producing a iron phosphate film which has a lower melting point than metal itself. This softer film causes a reduction in surface roughness and allows the load to be carried over a much bigger area. Some wear is created as a consequence but friction is lowered and with it, the temperature of contact. At higher loads (and temperatures) polyphosphate is formed and eventually thermal degradation of the phosphate takes place to release carbon. The latter is, of course, a good lubricant by virtue of its lamellar structure.

The above comments are, however, a simplification of the situation. If the iron salts that are formed are soluble in the oil they may desorb from the surface, leading to poor AW/EP performance. Interaction with other surface-active materials will inevitably influence the performance of AW/EP additives, while depletion in use due to oxidation, etc. will also affect performance.

Mechanism of Fire-Resistance of Hydraulic Fluids

Combustion is a vapor phase oxidation reaction and resistance to ignition depends on a number of factors that control the rate at which the vapor is first produced and then oxidized. These include:

(a) In the liquid phase:
- Fluid temperature increase:
- Fluid molecular weight/volatility:
- Heat of vaporization and:

(b) In the vapor phase:
- The temperature of the vapor:
- Thermal/oxidative stability of the vaporised fluid:
- The type and stability of radical intermediates formed:
- The heat released (or absorbed) when these radicals react and to what extent they react with radicals which would otherwise cause flame propagation.

Weil (1992) reports that, in the vapor phase, triphenyl phosphate breaks down into species of small molecular weight e.g., P_2, PO, PO_2 and HPO_2 and that these rapidly react to reduce the concentration of the hydrogen atoms (The reaction of the hydrogen atom with oxygen is said to be the rate controlling step in flame chemistry.).

$$H \cdot + O_2 \rightarrow HO \cdot + O$$

The combustion behavior of a fluid can be described in terms of (a) ignitability – or the temperature at which the fluid ignites under specific conditions, and (b) persistence of burning. If the fluid ignites then, when the ignition source has been removed, does the fluid propagate flame or does it self-extinguish?' Some products are more difficult to ignite but, once alight, continue to burn without the involvement of an external source of ignition.

Both these aspects need to be studied in order to get a complete picture of combustion behaviour and no single fire test has yet been able to provide an adequate overall picture.

Methods of Test

Antiwear and Extreme Pressure Additives

Attempting to simulate the exact conditions of use in a laboratory test in terms of test specimen geometry, loads, sliding speeds, temperature, presence of different metals etc. is difficult. Initially comparisons were made

under test conditions which could be readily replicated but which might not bear much relation to reality. For example the 4-Ball machine has been used for many years to compare the antiwear or extreme pressure properties of fluids and additives but without much relation to practice. However, it was eventually realized that in order to obtain relevant performance data it was necessary to test the fluids in the pumps, gears etc. that would be used in practice. As a consequence most modern circulatory oils and AW/EP additives are evaluated under standard test conditions which involve measuring the actual wear of pumps or the loads required to cause scuffing in gears. Table 1 lists the standard tests widely used for assessing the wear and load-carrying of circulatory oils while Table 2 shows typical antiwear and extreme pressure data on selected additives.

Ashless Phosphate Esters, Table 1 Common standard test procedures for assessing the AW/EP properties of circulatory oils

Test procedure	Antiwear properties	Extreme pressure properties
1. *General lab. methods*		
4-Ball method	ASTM D4172	ASTM D2783
Falex pin-on-block method	ASTM D2670	ASTM D3233
Timken machine	–	ASTM D2782
SRV machine	ASTM D6425	-
2. *Pump tests*		
Constant volume vane pump	ASTM D6943	–
	ASTM D7043	–
Piston pumps	ASTM D6813	–
3. *Gear tests*		
FZG gear test	–	ASTM D5182

Fire-Resistant Hydraulic Fluids

As with the laboratory representation of wear behaviour, it is also very difficult to simulate the widely varying conditions of combustion on a small scale, It is certainly not possible to assess the performance by evaluation under one test condition alone. Normally it is necessary to select tests representing the three main fire hazards e.g., ignition of a fluid jet or spray; ignition of a fluid falling on a hot surface and ignition when absorbed onto a substrate e.g., a wick test. Many different variations on these scenarios have been produced in the past and as a result, in order to ensure that products are compared under like conditions it has become necessary to standardize the test conditions. Conventional flash and fire points are no guides to the fire-resistance of higher molecular weight fluids and are not recommended as a means of differentiating fluid performance. However, it should be emphasized that although a phosphate is often described as fire-resistant it will ignite if heated to a high enough temperature for a long enough period of time.

Fire-resistance is also a relative term. As mineral oil is used so widely in industry, the combustion behaviour of a synthetic fluid is often compared with that of a mineral oil of similar viscosity. Table 3 shows the comparable behaviour of a triaryl phosphate and a mineral oil of the same viscosity grade under the above-mentioned test conditions (Phillips 2009b).

Ashless Phosphate Esters, Table 2 Typical antiwear/extreme pressure data on neutral and acid phosphates in oil and water

Test	*Wear scar diameter (mm)* – Additive at 2% in mineral oil			
	Mineral oil (Paraffinic)	Triaryl phosphate	Trialkyl phosphate	Ethoxylated acid phosphate
Four ball wear (ASTM D4172)	2.1	0.6	0.9	0.4
Falex EP test (ASTM D3233)	*Failure load (lbs)* – Additive at 2% in water (or oil)			
	Base – Water	Ethoxylated oleyl acid phosphate[a]	Ethoxylated 2-ethylhexyl phosphate[a]	Ethoxylated phenyl phosphate[a]
	<250	4,500	4,000	3,200
	Base – 100 N paraffinic oil	Ethoxylated oleyl acid phosphate	Ethoxylated cetyl-oleyl phosphate	Ethoxylated dinonylphenyl phosphate
	600	2,500	4,200	2,500

[a]Plus 2% triethanolamine

Ashless Phosphate Esters, Table 3 A comparison of the fire-test behaviour of mineral oil and a triaryl phosphate ester

Fluid (ISO VG 46)	Flash/Fire point ASTM D92 (°C)	Autoignition temperature ASTM D2155 (°C)	Manifold ignition temp. ISO 20823 (°C)	Spray ignition ISO/DIS 15029-2 (Ignitability class)
Mineral oil	250/285	390–400	370–395	H (worst)[a]
Triaryl phosphate	270–370	545–590	700–750	D–E

[a]The rating scheme currently specifies eight classes (or levels) of fire-resistance from A (the best) to H. Products containing a high water content would be expected to fall into Class A

Ashless Phosphate Esters, Table 4 Typical applications for phosphate esters as anti-wear/extreme pressure additives

Application	Triaryl phosphates	Trialkyl phosphates	Amine phosphates	Acid phosphates
Automotive				
ATF				x
Gear oil			x	
Power steering	x			
Shock absorber	x			
Industrial				
Hydraulic oils	x		x	
		x		
Gear oil			x	
Turbine oils	x	x		
Compressor oils	x			
Tractor oils	x			x
Metalworking	x	x	x	x
Grease	x		x	
Way oils	x			
Aviation				
Piston engine oil	x			
Turbine oil	x			
Grease	x		x	x

Key Applications

Antiwear and Extreme Pressure Additives

As was indicated above these additives are used in a variety of applications many of which are listed in Table 4 (Phillips 2011). In industrial oils the neutral phosphate is used at a concentration of 0.5–1.5% and, in addition to the antiwear performance, the following properties support their use in general industrial oils;

- Ashless
- Low odor, color and volatility
- Low acidity (unlikely to react with other additives) and non-corrosive.
- Good solubility for other additives
- Low level of toxicity
- Selected products are biodegradable
- Soluble in a wide range of basestocks

In metal working applications the acid phosphate content can be significantly higher (0.5–10%) depending on the severity of the application but levels are kept low where possible to minimise the potential for bacterial growth and an adverse effect on foaming. The following properties are advantageous for these additives when used in metalworking applications:

- Available as oil soluble, water soluble or emulsifiable products

- Low odor and color
- Activity can be "tailor made" to application
- Synergistic behaviour with sulfur
- Also provide excellent corrosion protection
- Stable under neutral and strong alkaline conditions

Fire-Resistant Hydraulic Fluids

These fluids are used where high temperatures or the absence of mechanical fire protection would preclude the use of mineral oil-based materials.

The chemistry and physical properties of the fluids largely determine their use. For example the alkyl phosphate esters have very good low temperature properties and are therefore used in aviation applications. On the other hand low temperature properties are not so important for terrestrial applications and the focus here is on fire-resistance. The following is a selection of general industrial applications:

Continuous casting	Hydraulic mining equipment
Die-casting	Fluid couplings
Hydraulic presses	Gate valve actuators
Hydraulic jigs	Air compressors
Injection moulding machines	Natural gas pumping
Oven controls	Hydraulic lifts e.g., on aircraft carriers
Automatic welders	Steam turbine control and lubrication
Billet loaders	Gas turbine control and lubrication
Furnace mechanisms	Grease base stock
Hot roll mills	Aircraft hydraulic fluid
Clay guns	Boiler controls
Ingot manipulators	
Foundry equipment	

Cross-References

► Additive Chemistry Testing Methods
► Aviation Turbine Engine Oil Application
► Elastohydrodynamic Lubrication
► Lubricant Formulation
► Steam Turbine Oils
► Tribochemistry of Extreme-Pressure Additives

References

F.T. Barcroft, S.G. Daniel, The action of neutral organic phosphates as EP additives, ASME. J. Basic Eng. 64-Lub-22 (1964)

E.R. Booser, M. Cohen, H. Lukas, Evaluation of fire-retardant fluids for turbine bearing lubricants, EPRI report NP-6542, Sep 1989

E.S. Forbes, H.B. Silver, The effect of chemical structure on the load-carrying properties of organo-phosphorus compounds. J. Inst. Pet. **56**(548), 90–98 (1970)

R.E. Hatton, Phosphate esters, in *Introduction to Hydraulic Fluids* (Reinhold, New York, 1962)

S.F. Jagger, A.N. Nicol, A.M. Thyer, A comprehensive approach to the assessment of fire-resistant hydraulic fluid safety, UK Health and Safety Laboratory Report, Health and Safety Executive, FR/02/05 (2005)

J.M. Kuchta, R.J. Cato, Ignitability and flammability properties of lubricants, SAE paper 680323, SAE (1968)

G. Livingstone, J. Prescott, D. Wooton, A lube oil that lasts forever. Turbomach. Int. Jan–Feb, 22–24 (2007)

D.H. Moreton, Development and testing of fire-resistant hydraulic fluids, SAE paper 490229, SAE, (1949)

W.D. Phillips, Turbine lubricating oils and hydraulic fluids, in *Fluids and Lubricants Handbook*, ed. by G.E. Totten, S.R. Westbrook, R.J. Shah (ASTM International, West Conshohocken, 2003)

W.D. Phillips, Ashless phosphorus-containing lubricating oil additives, in *Lubricant Additives – Chemistry and Applications*, ed. by L.R. Rudnick, 2nd edn. (CRC Press, Boca Raton, 2009a)

W.D. Phillips, Assessing and classifying the fire-resistance of industrial hydraulic fluids: the way ahead? J. ASTM Int. **6**, 6 (2009b)

W.D. Phillips, Phosphate esters, in *Handbook of Hydraulic Fluid Technology*, ed. by G.E. Totten, V.J. De Negri, 2nd edn. (CRC Press, Boca Raton, 2011)

W.D. Phillips, in The use of a fire-resistant turbine lubricant – Europe looks to the future, STLE STP 1407, ed. by W.R. Herguth, T.M. Warne. *Turbine Lubrication in the 21st Century*, STLE, West Conshohocken, PA, Jan 2001

W.D. Phillips, D.C. Placek, M.P. Marino, Neutral phosphate esters, in *Synthetics, Mineral oils and Bio-Based Lubricants*, ed. by L.R. Rudnick (CRC Press, Boca Raton, 2006)

D.G. Placek, S.G. Shankwalkar, Phosphate ester surface treatment for reduced wear and corrosion protection. Wear **173**, 207–217 (1994)

C.S. Saba, N.H. Forster, Further reactions of aromatic phosphate esters with metals and their oxides. Tribol. Lett. **12**(2), 135–146 (2002)

J.W.G. Staniewski, Maintenance practices for steam turbine control fire-resistant fluids. J. Syn. Lub. **23**, 109–135 (2006)

E.D. Weil, Phosphorus flame retardants, in *A Handbook of Phosphorus Chemistry*, ed. by R. Engel (Marcel Dekker, New York, 1992), pp. 683–739

G.F. Wolfe, M. Cohen, V.T. Dimitroff, Ten years experience with fire-resistant fluids in steam turbine electrohydraulic controls. Lubr. Eng. **25**(1), 6–14 (1970)

Asperities

Peter J. Blau
Materials Science and Technology Division,
Oak Ridge National Laboratory,
Oak Ridge, TN, USA

Synonyms

Contact spots; High spots; Protuberances

Definition

In tribology, asperities are high spots on surfaces that come into contact during wear or friction. The forms, shapes, and deformational characteristics of asperities play a role in modeling tribo-contacts, especially in cases where liquid-lubricating films are either absent or so thin that solid-to-solid interactions occur. Asperities that reside on contact surfaces can be visualized at a variety of size scales, depending upon the phenomena being considered. The fracture, deformation, and/or loss of asperities is associated with wear, polishing, and either the roughening or smoothing of surfaces.

Scientific Applications

The concept of an asperity is fundamental to the visualization and modeling of friction and wear-related phenomena; especially in situations where solid-to-solid interactions occur. Simply stated, an asperity is a high spot (bump) on a surface that either makes contact with an opposing surface or affects the flow of material in a lubricating film or mass of entrained particles. The material of which an asperity is composed can deform elastically, plastically, and/or visco-elastically. The number of asperities that reside on a surface can vary from one (e.g., a cluster of atoms on the sharp probe tip of an atomic forces microscope) to many millions (comprising the surface finish of a typical engineering bearing component). In addition, the number and shapes of asperities that support the applied load can change during wear or frictional contact. Thus, the principal attributes, including the lifetime, of a given asperity depend upon tribological conditions and the material of which the asperity is composed.

In principle, asperities can be as small as individual atoms that nestle next to each other on a contacting solid surface, but at the larger sizes, they can be meters in extent, like the highly-stressed, smeared interfaces within a geological fault lying deep within the earth. Therefore, it is important to define the scale of the interacting asperities for the tribosystem of interest. The terms "micro-asperities" and "nano-asperities" have been used in some instances to differentiate fine-scale bumps from larger protuberances on a surface. In that sense, there can be small asperities on top of larger asperities. Therefore, the context of usage of the term asperity is not without ambiguity, and the term should be defined for the situation under consideration.

Not only do the numbers and sizes of asperities vary between tribosystems, but they also assume many complex shapes, depending to a great extent on the manner in which the surfaces were created. Prior to the advent of modern microscopes and profiling instruments, early philosophers who contemplated the origins of friction envisioned asperities as spheres, cones, or interlocking saw-teeth (Dowson 1998). Contemporary thinking about asperities was for many years influenced by the work of Archard (1953) and Holm (1967). For decades during the mid-twentieth century, models for friction and wear (especially abrasive wear) were based upon such visualizations. In sliding wear studies, simplifying assumptions commonly involved defining a surface as a set of hemispherical asperities with a normal distribution of radii and/or heights (e.g., Greenwood and Williamson 1966). In other approaches, hard conical asperities were assumed to have a fixed apex angle as they plowed or cut through a surface (e.g., Hokkirigawa and Kato 1988). Commonly, early models for asperities limited the deformation of the interfacial material to only one side of the contact pair, assuming that the opposite counterface was perfectly rigid, perfectly flat, sharp and non-wearing, and/or non-deformable. Eventually, however, asperity-based models grew increasingly complex, and high-performance computers could be used to calculate the deformation of a set of individual asperities within a region of sliding contact. Therefore, the actual shapes of evolving contact surface features, like those observed in optical or electron microscopes, could more closely be modeled (e.g., Liu et al. 2003; Kim et al. 2007).

When two surfaces comprised of asperities first touch, they begin contacting at the highest points. At first contact, the local pressure is extremely high because only a minute area supports the normal force (load). Depending on the magnitude of the normal force, the tallest asperities deform until there is sufficient bearing area developed to support that load. The intermeshing of surfaces in a tribo-contact therefore involves engaging two sets of asperities, one upon the other. Even if the surfaces are of similar roughness, a small relative rotation or change in the orientation ("lay") of the two solid surfaces can result in different numbers of asperity interactions and contact geometry. Vibration during sliding contact can also change the instantaneous distribution of asperity contacts. The process of wear can result in *asperity truncation* in which the high spots are removed, leaving a new distribution of heights and shapes of asperities. Valleys between load-bearing asperities in liquid-lubricated systems are sometimes envisioned as reservoirs that serve to retain lubricant.

An intentional process of running-in is commonly employed to smooth initially sharp asperities on fresh bearing surfaces. In metallic bearings, the process increases interfacial conformity and leads to more

efficient fluid-film lubrication. In terms of the ratio of lubricant film thickness to root-mean-square roughness (conventionally designated as the λ ratio), running-in reduces roughness (the denominator) and hence tends to increase λ, which, in turn, can lead to lower friction and a more favorable lubrication regime (see "► Stribeck Curves"). During dry or severe sliding contact, original surface asperities can be submerged under films of transferred material or wear debris, leading to a new load-bearing condition.

Cross-References

► Adhesive Contact of Elastic Bodies
► Adhesive Contact of Inelastic Bodies
► A-Spots
► Contacts Considering Adhesion
► Contact of Rough Surfaces
► Friction (Concepts)
► Friction, History of Research
► Microtribodynamics of Magnetic Storage Hard Disk Drives
► Topography of Engineering Surfaces

References

J.F. Archard, Contact and rubbing of flat surfaces. J. Appl. Phys. **24**(8), 981–988 (1953)

F.P. Bowden, D. Tabor, *Friction and Lubrication of Solids* (Oxford University Press, Oxford, 1996), 337 pp. Oxford Press, originally published 1950 this was republished in 1996 Oxford University Classic Texts in the Physical Sciences series

D. Dowson, *History of Tribology*, 2nd edn. (Wiley, New York, 1998), 768 pp

J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. Proc. R. Soc. Lond. **A295**, 300–319 (1966)

H. Hokkirigawa, K. Kato, An experimental and theoretical investigation of ploughing, cutting, and wedge formation during abrasive wear. Tribol. Int. **21**(1), 51–57 (1988)

R. Holm, *Electric Contacts – Theory and Applications*, 4th edn. (Springer, New York, 1999), 516 pp. (First published by Springer, Berlin, 1967)

H.J. Kim, W.K. Kim, M.L. Falk, D.A. Rigney, MD simulations of microstructure evolution during high-velocity sliding between crystalline materials. Tribol. Lett. **28**, 299–306 (2007)

G. Liu, Q. Wang, S. Liu, A three-dimensional thermal-mechanical asperity contact model for two nominally flat surfaces in contact. J. Tribol. **123**(3), 595–563 (2003)

E. Rabinowicz, *Friction and Wear of Materials*, 2nd edn. (Wiley, New York, 1995), 336 pp

Asperity Contact

► Contact of Rough Surfaces: The Greenwood and Williamson/Tripp, Fuller and Tabor Theories

Asperity Contact Theories

► Contact of Rough Surfaces: The Greenwood and Williamson/Tripp, Fuller and Tabor Theories

A-Spots

Peter J. Blau
Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Definition

In tribology, *a-spots* are tiny areas on the surfaces of contacting solid bodies through which electric current flows under an applied potential. Conduction occurs in regions that lie within the boundaries of asperity contacts, at individual locations or at clusters of conductive areas. This term is associated with the work of Ragnar Holm in the mid-20th century. A-spots are not equivalent to asperities.

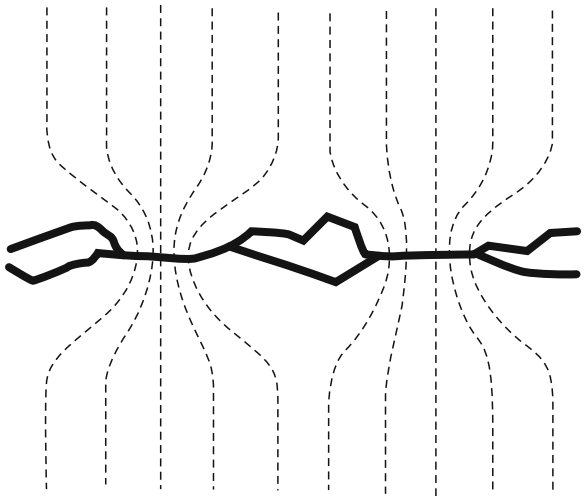
Scientific Fundamentals

A central concept in tribology is that the apparent area of surfaces in contact is larger than the actual contact area. The actual ("true") contact area is the sum of load-bearing asperity contacts. Conductive a-spots lie within the boundaries of individual asperity contacts. Current can flow through breeches in insulating layers, like oxide films or adsorbed species that may be ruptured as a result of mechanical contact. Current can also flow if tunneling occurs through thin surface films. The localization of current flow between rough, contacting surfaces gives rise to a phenomenon known as *constriction resistance*, as shown in Fig. 1.

Electrical resistance (R), expressed in ohms (Ω), is directly proportional to a material's resistivity (ρ), expressed in units of Ω/m; the length of the element through which the current passes (L); and inversely proportional to the area of the element (A) perpendicular to the direction of current flow. Thus,

$$R = \frac{\rho L}{A} \tag{1}$$

In a tribocontact, the localized contact area can be comprised either of individual a-spots or clusters of a-spots.



A-Spots, Fig. 1 A-spots conduct electrical current (represented with *dashed lines*) within asperity contacts, and this results in constriction resistance

Holm (1967) expressed the resistance of a single, film-covered asperity between two bodies having bulk resistivities ρ_1 and ρ_2, respectively, as follows:

$$R = \left(\frac{\rho_1 + \rho_2}{4a}\right) + \left(\frac{\sigma}{\pi a^2}\right) \quad (2)$$

in which the conducting asperity radius is a and σ represents resistance per unit area of the surface film. Based on this simplified expression, the total constriction resistance can be determined by summing the resistances of individual a-spots. Holm's early model has since been extended and refined beyond these basic concepts.

The number and sizes of a-spots that form with an apparent area of contact is related to the applied load. When surfaces are brought together under load, contact between the tallest asperities occurs first. As load is increased, there is localized deformation and the generation of a static bearing area that is proportional to the applied load. This concept is analogous to that of indentation hardness. The relationship of R to increasing load in metal-metal electric contacts, like those involving gold, rhodium, and noble metal alloys, is non-linear, and reflects the influence of surface films, adsorbed species, contaminants, and other factors. Discussions of these phenomena may be found in the books by Holm (1999), Slade (1999), and Braunovic et al. (2006).

The concept of a-spots has been attributed to groundbreaking research on electrical contacts, motor brushes, and mechanical relays conducted by Ragnar Holm in the mid-1900s, and in 1953, the Institute of Electrical and Electronics Engineers (IEEE) established a conference on electrical contact phenomena in his honor. Holm's studies of a-spots influenced the work of well-known tribophysicists like J. F. Archard (1953) and J. A. Greenwood and J. B. P. Williamson (1966), who based models for friction and wear on surface geometry.

Cross-References

► Asperities
► Electric Contact, Elements, and Systems

References

J.F. Archard, Contact and rubbing of flat surfaces. J. Appl. Phys. **24**(8), 981–988 (1953)

M. Braunovic, V.V. Konchits, N.K. Myshkin, *Electrical Contacts – Fundamentals, Applications, and Technology* (CRC Press, Boca Raton, 2006). 645 pp

J.A. Greenwood, J.B.P. Williamson, Contact of nominally flat surfaces. Proc. R. Soc. Lond. **A295**, 300–319 (1966)

R. Holm, *Electric Contacts – Theory and Applications*, 4th edn. (Springer, New York, 1999). 516 pp. (First published by Springer, Berlin, 1967)

P.G. Slade (ed.), *Electrical Contacts – Principles and Applications* (CRC Press, Boca Raton, 1999). 1073 pp

Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts

Ilya I. Kudish
Department of Mathematics, Kettering University, Flint, MI, USA

Synonyms

Methods of matched asymptotic expansions in heavily loaded EHL contacts; Perturbation methods in heavily loaded EHL contacts

Definition

Asymptotic methods in heavily loaded EHL contacts are a series of specialized perturbation techniques capable of producing accurate analytical results. The asymptotic methods are based on competing physical mechanisms occurring in heavily loaded EHL contacts that result in the presence of a small parameter in EHL problems. Application of these methods produces problem solutions in the form of asymptotic expansions in this small parameter.

Scientific Fundamentals

Heavily loaded EHL contacts can be found in various moving machine elements such as bearings and gears. In heavily loaded lubricated contacts at least one of the following conditions takes place: the load applied to the contact is high, the contact region is small, contact surfaces are involved in slow motion, or the lubricant viscosity is low. Under such conditions the lubrication film thickness is small and the contact pressure in the central part of the contact is close to the pressure in a dry contact of the same contact surfaces (Hertzian pressure). In a lubricated contact the pressure departs from the Hertzian one only in narrow zones at the entrance to and exit from the lubricated contact. In spite of its narrowness the inlet zone is very important as it is the region where the lubrication film thickness is formed. In the inlet and exit zones both elastic and fluid effects are equally important while in the central region elastic effects dominate. Such solution features are typical for problems with boundary layers in which solution experiences a fast transition from one set of properties and mechanisms controlling them to another set of properties and mechanisms. It is typical for such problems to have a small parameter involved in the governing equations. That offers an opportunity to consider application of certain asymptotic methods for solution of these problems.

In the classic formulation the EHL problem has been studied in numerous papers and monographs by numerical and approximate analytical methods for over 60 years. Some references to these papers and monographs can be found in Dowson and Higginson (1966), Ertel (1945), and Grubin (1949); the latest achievements in numerical methods of solving the considered problem are described by Houpert and Hamrock (1986), Bissett and Glander (1988), Hamrock et al. (1988) (also see the bibliography in these papers), Venner and Lubrecht (2000), Evans and Hughes (2000), and others.

The lubrication regimes considered here correspond to the case of a relatively weak ("non-prevailing" in any zone of a contact) dependence of lubricant viscosity μ on pressure p. Other types of lubrication regimes are considered in Kudish and Covitch (2010). These regimes were studied earlier in Dowson and Higginson (1966), Houpert and Hamrock (1986), Bissett and Glander (1988), Hamrock et al. (1988), Venner and Lubrecht (2000), and Evans and Hughes (2000) using numerical methods. Ertel (1945), Grubin (1949), Archard and Baglin (1986), and Crook (1961) considered cases of rapidly growing lubricant viscosity with pressure using approximate analytical methods. The method used by Crook (1961) differs from the methods employed by Ertel (1945) and Grubin (1949) only by more precise techniques under the same prior assumptions. The method proposed by Archard et al. (Archard and Baglin 1986) is an extension of the methods published in Ertel (1945), Grubin (1949), and Crook (1961) for the case of a weak relationship between the viscosity μ and pressure p. The main difference between the methods published in Ertel (1945), Grubin (1949), Archard and Baglin (1986), and Crook (1961) is in the approximations of gap function $h(x)$ and in a parabolic approximation for pressure $p(x)$ in the zone of large pressure.

The main mathematical difficulties in the analytical and numerical analysis of the classic EHL problem for line contacts are:

1. The essential nonlinearity of the problem, causing the possibility of the existence of solutions with qualitatively and quantitatively different structure depending on the values of the problem input parameters.
2. The integro-differential form of the problem equations, causing the existence of a small parameter ω and boundary layers for heavily loaded lubrication regimes.
3. The proximity of the considered problem to the classic contact problem of elasticity described by an integral equation of the first kind, numerical solutions of which are generally unstable. (The latter problem is the limiting case for the EHL problem for the small parameter $\omega = 0$.)
4. The presence of the unknown dimensionless free exit boundary $x = c$ (exit point) and the dimensionless exit lubrication film thickness H_0.
5. The possibility to discriminate and differentiate between the influence of different (inlet and exit) zones and the Hertzian region of the lubricated contact on the EHL problem solution.

Historically, the problems of the elastohydrodynamic lubrication (EHL) theory were predominantly studied by direct numerical methods. Due to the complexity and ill-conditioned nature of EHL problems for heavily loaded contacts, the precision and stability of such numerical solutions deteriorates as the load conditions get heavier. As was mentioned earlier, there are a few papers that explore approximate analytical approaches to solution of EHL problems. These analytical approaches all use certain assumptions about the problem solutions that may not adequately represent the solution features. An alternative to direct numerical solution of EHL problem – the asymptotic approach to solution of EHL problems for heavily loaded contacts – is provided here. The asymptotic approach provides the opportunity to obtain some

analytical results followed by numerical solution of the asymptotic equations. The benefits of the asymptotic approach are some analytical results (e.g., formulas for the lubrication film thickness), clear structure and characteristic sizes of different regions in lubricated contacts, reduction of the number of the problem input parameters, and a distinct way to propose a regularization approach to make numerical solutions of the asymptotic equations as well as of the original EHL problems for heavily loaded contacts stable.

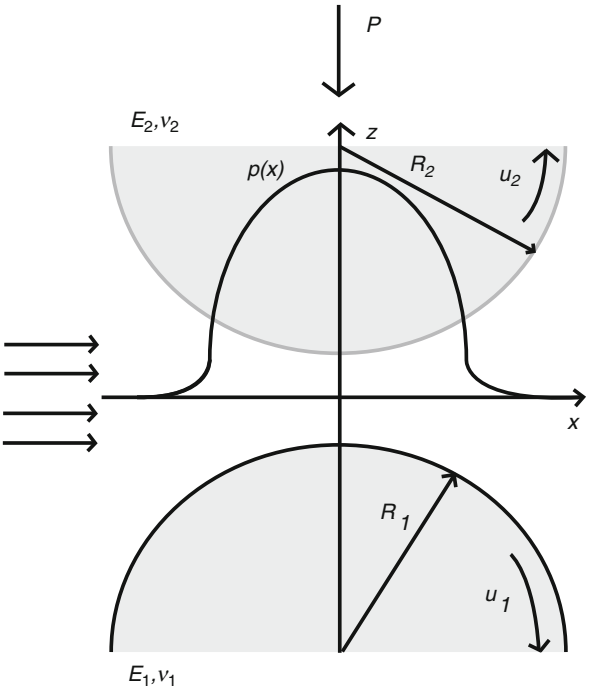
Numerical solutions of obtained asymptotic equations not only provide validation of the asymptotic analysis in cases of precritical lubrication regimes but, under certain conditions, also show their validity for cases of over-critical lubrication regimes (Kudish and Covitch 2010).

The purpose here is to give a complete analysis of the aforementioned classic problem without any contradictions and prior assumptions about its solution as well as to understand better the structure of the EHL problem solution.

A method of matched asymptotic expansions (Van-Dyke 1964) is applied to study the equations of the EHL problem. The first time the method of matched asymptotic expansions was applied to the simplest lubrication problem with Newtonian fluid in an isothermal heavily loaded contact was in Kudish (1976, 1982). In Kudish (1978a, b) this method has been further developed for the case of non-Newtonian fluids under iso- and non-isothermal conditions using the method of regular and matched asymptotic expansions (Van-Dyke 1964). Also, in Kudish (1978a, b, 1983) the conditions for the existence of the second pressure spike and for the dependence of its magnitude, width, and location on the inlet coordinate, temperature, and so on have been established. In all these cases, the structure of the contact region and the solution behavior were studied and asymptotically valid formulas for the film thickness were obtained.

Equivalent Formulations of EHL Problems

Consider a steady-plane isothermal EHL problem for a heavily loaded contact of two moving infinite elastic cylinders with smooth surfaces with velocities u_1 and u_2 (see Fig. 1). The cylinders' radii are R_1 and R_2 while the elastic moduli and Poisson's ratios of the cylinders' materials are E_1 and E_2 and ν_1 and ν_2, respectively. The lubricant is considered to be an incompressible Newtonian fluid with the viscosity μ exponentially dependent on pressure $\mu = \mu_a \exp(\alpha_p p)$, where μ_a is the ambient lubricant viscosity, α_p is the pressure viscosity coefficient, and p is



Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts, Fig. 1 General view of a lubricated contact

the lubricant pressure. The cylinders are loaded with normal load P per unit length. Assume that the lubrication film thickness/gap h is much smaller than the size of the contact $x_e - x_i$, which, in turn, is much smaller than the cylinders' radii R_1 and R_2 (here x_i and x_e are the inlet and exit points of the contact). Using the dimensionless variables

$$x' = \frac{x}{a_H}, \quad a = \frac{x_i}{a_H}, \quad c = \frac{x_e}{a_H}, \quad p' = \frac{p}{p_H}, \quad h' = \frac{h}{h_e}, \quad \mu' = \frac{\mu}{\mu_a} \tag{1}$$

and parameters

$$V = \frac{24\mu_a(u_1 + u_2)R'^2}{a_H^3 p_H}, \quad Q = \alpha_p p_H, \quad H_0 = \frac{2R'h_e}{a_H^2}, \tag{2}$$

the EHL problem equations are reduced to (primes at dimensionless variables are omitted)

$$\frac{d}{dx}\left[\frac{H_0^2}{V}\frac{h^3}{\mu}\frac{dp}{dx} - h\right] = 0, \tag{3}$$

$$p(a) = p(c) = \frac{dp(c)}{dx} = 0, \tag{4}$$

$$H_0(h-1) = x^2 - c^2 + \frac{2}{\pi}\int_a^c p(t)\ln\left|\frac{c-t}{x-t}\right|dt, \quad (5)$$

$$\int_a^c p(t)dt = \frac{\pi}{2}, \quad (6)$$

where p_H and a_H are the maximum Hertzian pressure $\left(p_H = \sqrt{\frac{E'P}{\pi R'}}\right)$ and the Hertzian contact semi-width $\left(a_H = 2\sqrt{\frac{R'P}{\pi E'}}\right)$, respectively. In the above equations the dimensionless lubricant viscosity is taken according to an exponential law in the form

$$\mu = \mu(p, Q) = \exp(Qp), \quad (7)$$

where Q is the dimensionless pressure viscosity coefficient.

In (3)–(6) the inlet coordinate a, the speed-load parameter V, and the pressure coefficient of viscosity Q are considered to be known while the exit coordinate c and the lubrication film thickness H_0 together with the functions of pressure $p(x)$ and gap $h(x)$ in the contact region (a, c) have to be determined from the solution of the problem.

Also, the problem (3)–(6) can be presented in another equivalent form. Solving (5) for pressure $p(x)$ and taking into account boundary conditions (4) the equivalent form (Vorovich et al. 1974) of the system can be expressed in the form

$$p(x) = R(x)\left[1 - \frac{1}{2\pi}\int_a^c \frac{dM(p,h)}{dt}\frac{dt}{R(t)(t-x)}\right], \quad (8)$$
$$R(x) = \sqrt{(x-a)(c-x)},$$

$$\int_a^c \frac{dM(p,h)}{dt}\frac{dt}{R(t)} = \pi(a+c), \quad \int_a^c \frac{dM(p,h)}{dt}\frac{tdt}{R(t)} = \pi\left[\left(\frac{c-a}{2}\right)^2 + \frac{(a+c)^2}{2} - 1\right], \quad (9)$$

$$M(p,h) = \frac{H_0^3}{V}\frac{h^3}{\mu}\frac{dp}{dx}, \quad (10)$$

$$H_0(h-1) = x^2 - c^2 + \frac{2}{\pi}\int_a^c p(t) \quad \ln\left|\frac{c-t}{x-t}\right| dt. \quad (11)$$

It is obtained by inverting the singular Cauchy integral in (3)–(5). The equivalence of the systems of (3)–(6) and (8)–(11) takes place under the following condition (Vorovich et al. 1974):

$$\frac{1}{\pi}\int_a^c M(p,h)\frac{dt}{R(t)} + \frac{2}{\pi}\int_a^c p(t)\ln\frac{1}{|c-t|}dt + c^2 - \frac{1}{2}\left(\frac{c-a}{2}\right)^2 - \left(\frac{a+c}{2}\right)^2 = \ln\left|\frac{4}{c-a}\right|. \quad (12)$$

Moreover, due to the equivalence of systems (3)–(5) and (8)–(11) in the latter system, (11) for gap $h(x)$ can be replaced by the (Reynolds) equation

$$M(p,h) = H_0(h-1), \quad (13)$$

which follows from integration of (3) with the last boundary condition from (4) (also see (10)). Therefore, system of (8)–(11) is equivalent to the system of (8)–(10) and (13).

Asymptotic Analysis of a Heavily Loaded Precritical Lubrication Regime

The lubrication regime is called heavily loaded if in the Hertzian region (contact region of dry solids) that is away from contact boundaries (determined by inequalities $x - a \gg \varepsilon_q$ and $c - x \gg \varepsilon_g$) the following estimates:

$$H_0(h-1) \ll 1 \quad \text{and} \quad \frac{dM(p,h)}{dx} \ll 1 \quad \text{for} \quad x - a \gg \varepsilon_q \quad \text{and} \quad c - x \gg \varepsilon_g, \quad (14)$$

are valid. The characteristic size of the inlet zone $\varepsilon_q(\omega)$ is determined by the given inlet coordinate

$$a = -1 + \alpha_1\varepsilon_q, \quad \alpha_1 = O(1), \quad \omega \ll 1, \quad (15)$$

while the exit coordinate c and the characteristic size of the exit zone $\varepsilon_g(\omega)$ are found from

$$c = 1 + \beta_1\varepsilon_g + o(\varepsilon_g), \quad \beta_1 = O(1), \quad \omega \ll 1. \quad (16)$$

Here ω is a small parameter of the problem (depending on the problem parameters $\omega = V \ll 1$ or $\omega = Q^{-1} \ll 1$). Using the definition of a heavily loaded contact (14), from (8) it can be found in the external region that

$$p(x) = p_0(x) + o(1), \quad p_0(x) = \sqrt{(x-a)(c-x)}, \quad x - a \gg \varepsilon_q \text{ and } c - x \gg \varepsilon_g. \quad (17)$$

Taking the coordinate of the exit point c according to (16) and using the assumption that $\varepsilon_q = O(\varepsilon_g)$, the following estimates $p_0(x) = O\left(\varepsilon_q^{1/2}\right)$ for $r = (x-a)/\varepsilon_q = O(1)$ and $s = (x-c)/\varepsilon_q = O(1)$,

$\omega \ll 1$, are obtained. As a result, in the inlet and exit zones the solution is searched in the form

$$p(x) = \varepsilon_q^{1/2} q(r) + o\left(\varepsilon_q^{1/2}\right), \quad q(r) = O(1), \quad r = O(1), \quad w \ll 1, \tag{18}$$

$$p(x) = \varepsilon_g^{1/2} g(s) + o\left(\varepsilon_g^{1/2}\right), \quad g(s) = O(1), \quad s = O(1), \quad w \ll 1, \tag{19}$$

where $q(r)$ and $g(s)$ are the main terms of pressure $p(x)$ asymptotic expansions in the inlet and exit zones.

Consider (8) in the inlet zone. The integral in (8) can be expressed as a sum of tree integrals: over the inlet and exit zones and over the Hertzian region. By using expressions (18), (19), and estimates

$$\mu(p, Q) = O(1) \text{ for } r = O(1) \text{ and } s = O(1), \quad \omega \ll 1 \tag{20}$$

for this class of relations for $\mu(p, Q)$, the main term of the asymptotic for the lubrication film thickness H_0 becomes (Kudish and Covitch 2010; Kudish 1976)

$$H_0 = A\left(V\varepsilon_q^2\right)^{1/3} + \ldots, \quad \omega \ll 1, \tag{21}$$

where $A(\alpha_1) = O(1)$ for $\omega \ll 1$ and $A(\alpha_1)$ is an unknown non-negative constant independent from ω and ε_q. The formula (21) is obtained by taking into account the fact that the elastic and fluid flow contributions to the problem solution in the inlet zone are of the same order of magnitude. Based on the estimate (21) a similar analysis of the exit zone allows us to set

$$\varepsilon_g = \varepsilon_q, \tag{22}$$

and to derive the following asymptotically valid equations for $q(r)$ in the inlet and exit zones (Kudish and Covitch 2010):

$$q(r) = \sqrt{2r}\left[1 - \frac{1}{2\pi}\int_0^\infty \frac{d}{dt} M_0(q,\ h_q,\ \mu_q,\ t)\frac{dt}{\sqrt{2t}(t-r)}\right], \tag{23}$$

$$g(s) = \sqrt{-2s}\left[1 - \frac{1}{2\pi}\int_{-\infty}^0 \frac{d}{dt} M_0(g, h_g, \mu_g, t)\frac{dt}{\sqrt{-2t}(t-s)}\right], \tag{24}$$

where functions $h_q(r)$, $h_g(s)$, $\mu_q(q)$, and $\mu_g(g)$ are the main terms of the asymptotic expansions of $h(x)$ and $\mu(p, Q)$ in the inlet and exit zones, respectively.

Using (14), (15), (16), (18), (19), (21), and (22), the asymptotic analysis of (9) and (10) leads to a system of linear algebraic equations for A^3 and β_1. The solution of this system has the form (Kudish and Covitch 2010)

$$1 = \pi\alpha_1 \Big/ \int_0^\infty \frac{d}{dt} M_0(q, h_q, \mu_q, t)\frac{dt}{\sqrt{2t}}, \tag{25}$$

$$\beta_1 = \frac{1}{\pi}\int_{-\infty}^0 \frac{d}{dt} M_0(g, h_g, \mu_g, t)\frac{dt}{\sqrt{-2t}}. \tag{26}$$

Also, (25) can be represented in the form

$$A^3 = \pi\alpha_1 \Big/ \int_0^\infty \frac{d}{dt}\left(\frac{h_q^3}{\mu_q}\frac{dq}{dt}\right)\frac{dt}{\sqrt{2t}}. \tag{27}$$

Now, it is useful to introduce the definitions of the regimes of starved and fully flooded lubrication that are determined by the lubricant flux entering the contact or the position of the inlet coordinate a. These lubrication regimes can be easily expressed by means of the order of proximity of the inlet coordinate a to the left boundary $x = -1$ of the Hertzian dry contact. The conditions when

$$h(x) - 1 \ll 1 \text{ for all } x \in [a, c], \quad \omega \ll 1 \tag{28}$$

are called regimes of starved lubrication, and the conditions when

$$h(x) - 1 = O(1) \text{ for } x - a = O(\varepsilon_q) \text{ and } c - x = O(\varepsilon_q), \quad \omega \ll 1 \tag{29}$$

are called regimes of fully flooded lubrication. From (21), (22), and the estimates of the integrals for $h(x)$ in the inlet and exit zones follows that starved lubrication regimes (28) take place under the condition

$$\varepsilon_q^{3/2} \ll H_0, \quad \omega \ll 1, \tag{30}$$

and fully flooded lubrication regimes (29) under the condition

$$\varepsilon_q^{3/2} = O(H_0), \quad \omega \ll 1. \tag{31}$$

Thus, using formula (21) for the exit film thickness H_0 the conditions for starved

$$\varepsilon_q \ll V^{2/5}, \quad \omega \ll 1, \tag{32}$$

and fully flooded

$$\varepsilon_q = V^{2/5}, \quad \omega \ll 1, \tag{33}$$

lubrication regimes are obtained. For fully flooded lubrication regimes (21) and (33) lead to the formula for the lubrication film thickness

$$H_0 = AV^{3/5}, \quad \omega \ll 1. \tag{34}$$

Finally, using (21), (32), and (33) the final equations for starved and fully flooded lubrication regimes can be written in the form, respectively

$$h_q(r) = 1, \tag{35}$$

$$h_g(s) = 1; \tag{36}$$

$$A\,[h_q(r) - 1] = \frac{2}{\pi}\int_0^\infty [q(t) - q_a(t)]\ \ln\frac{1}{|r-1|}\,dt, \tag{37}$$

$$A\,[h_g(s) - 1] = \frac{2}{\pi}\int_{-\infty}^0 [g(t) - g_a(t)]\ \ln\frac{1}{|s-t|}\,dt. \tag{38}$$

However, the last two equations can be replaced by the equivalent Reynolds equations in the inlet and exit zones

$$\begin{aligned} &A\,[h_q(r) - 1] = M_0(q,\ h_q, \mu_q, r)\quad if\ r = O(1);\\ &A[h_g(s) - 1] = M_0(g, h_g, \mu_g, s)\quad if\ s = O(1),\\ &M_0(p,\ h, \mu, x) = A^3\frac{h^3}{\mu}\frac{dp}{dx}, \end{aligned} \tag{39}$$

where M_0 is the main term of the asymptotic expansions of function M in the inlet and exit zones.

Therefore, in the inlet zone the problem is reduced to the system of (23), (27), and (35) or (37) (depending on whether the lubrication regime is starved or fully flooded) for functions $q(r)$, $h_q(r)$, and constant A. In the exit zone the problem is reduced to the system of (24), (26) and (36) or (38) (depending on whether the lubrication regime is starved or fully flooded) for functions $g(s)$, $h_g(s)$, and constant β_1.

It should be pointed out that asymptotic relationships $q(r) \to q_a(r) = \sqrt{2r} + \alpha_1/\sqrt{2r},\ r \to \infty$ and $g(s) \to g_a(s) = \sqrt{-2s} - \beta_1/\sqrt{-2s},\ s \to -\infty$ can be obtained from (15), (16), (23)–(26). Obviously, the accepted assumption (14) is valid.

It can be shown that for $\mu = 1$ and regimes of starved lubrication the solutions of the asymptotically valid systems of equations in the inlet and exit zones have the following properties

$$\begin{aligned} &q(r,\alpha_1) = |\alpha_1|^{1/2} q(r/|\alpha_1|, -1),\ g(s,\alpha_1) = |\alpha_1|^{1/2} g(s/|\alpha_1|, -1),\\ &A(\alpha_1) = |\alpha_1|^{2/3}\theta(-\alpha_1)A(-1),\ \beta_1(\alpha_1) = |\alpha_1|\beta_1(-1), \end{aligned} \tag{40}$$

where $\theta(x)$ is a step function, $\theta(x) = 0$, $x \le 0$ and $\theta(x) = 1$, $x > 0$. It follows from (27) that the formulation of the problem makes sense only for $a \le -1$ ($\alpha_1 \le 0$), which means that $H_0 \ge 0$ ($A \ge 0$). Obviously, $H_0 = A = 0$ for $a = -1$ ($\alpha_1 = 0$).

After the asymptotic solutions of the problem in the inlet and exit zones are obtained one can determine the uniformly valid approximate solution of the problem for pressure $p_u(x)$ and gap $h_u(x)$ in the form (Van-Dyke 1964)

$$\begin{aligned} p_u(x) &= \frac{\varepsilon_q}{2}\,q\left(\frac{x-a}{\varepsilon_q}\right)g\left(\frac{x-c}{\varepsilon_q}\right),\\ h_u(x) &= h_q\left(\frac{x-a}{\varepsilon_q}\right)h_g\left(\frac{x-c}{\varepsilon_q}\right). \end{aligned} \tag{41}$$

where a and c are determined by formulas (15) and (16).

The solutions of the asymptotic systems depend on a smaller number of the input problem parameters than the solution of the original problem (3)–(7).

Estimates (32) and (33) define the so-called precritical lubrication regimes. In general, for a given $\varepsilon_q = \varepsilon_q(V, Q)$, the precritical regimes are characterized by the following condition:

$$\mu\left(\varepsilon_q^{1/2},\ Q\right) = O(1),\ \ \varepsilon_q \ll 1,\ \ \omega \ll 1, \tag{42}$$

while the over-critical regimes are determined by the condition

$$\mu\left(\varepsilon_q^{1/2},\ Q\right) \gg 1,\ \ \varepsilon_q \ll 1,\ \ \omega \ll 1. \tag{43}$$

The detailed treatment of over-critical lubrication regimes is given in Kudish and Covitch (2010).

The above asymptotic methodology can be successfully applied to the analysis of heavily loaded contacts lubricated by fluids with non-Newtonian rheology (see ▶ Thermoelastohydrodynamically Lubricated Contacts with Non-Newtonian Lubricants: Asymptotic Approach; Kudish and Covitch 2010).

Numerical Validation of the Asymptotic Analysis of EHL Problems

A detailed numerical analysis of the asymptotic and original EHL problems is made in Thermal EHL Contacts with Non-Isothermal Lubricants: Asymptotic Approach, ▶ Effect of Starvation on EHL Film Thickness, ▶ Numerical Stability and Precision in Elastohydrodynamic Lubrication (EHL) and in Kudish and Covitch (2010).

To validate the asymptotic approach developed earlier for precritical lubrication regimes it is necessary to compare the solutions of the EHL problem in the asymptotic and original formulations. Generally, it can be done in two different ways: on either the conceptual or detailed numerical level.

On the conceptual level it can be done as follows. If the asymptotic approach is valid then for fully flooded precritical lubrication regimes the formula for the film thickness $H_0 = AV^{3/5}$ (see formula (34)) should be correct. More specifically, for large enough $|a|$, $a < 0$ (which can be judged by the value of $h(a)$ in comparison with 1), the values of coefficient A and $\beta_1 = (c - 1)/V^{2/5}$ (see formulas (16) and (33)) are supposed to be functions of only $Q_0 = QV^{1/5}$. For the asymptotic method to be valid, the characteristic size of the inlet zone $\varepsilon_q = V^{2/5}$ (see formula (33)) should be small and the regime of lubrication is supposed to be precritical. Therefore, for sufficiently large $|a|$, $a < 0$, and practically fixed value of parameter $Q_0 = QV^{1/5}$ (because Q_0 is fixed while V is small) for different values of parameter V, the value of coefficient $A = H_0V^{-3/5}$ should be practically constant. Consider the following two examples. For $V = 0.05$ and $V = 0.1$ the values of $\varepsilon_q = 0.302$ and $\varepsilon_q = 0.398$ are relatively small. For $Q = 0$ equation $\mu\left(\varepsilon_q^{1/2},\ Q\right) = 1$ means that the lubrication regime is precritical (see definition of precritical regime (42)). Notice that the solution of the original (not asymptotic) EHL problem for $Q = 0$ and $a = -2$, $V = 0.05$, gives $H_0 = 0.067$ and $c = 1.024$, while for $a = -2$, $V = 0.1$ it gives $H_0 = 0.102$ and $c = 1.030$. Therefore, for $a = -2$, $Q = 0$, $V = 0.05$, and $V = 0.1$ the values of coefficient $A = H_0V^{-3/5}$ are equal to 0.403 and 0.407, while the values of $\beta_1 = (c - 1)/V^{2/5}$ are equal to 0.078 and 0.075, respectively. The differences between these pairs of values of A and β_1 are about 1% and 4%. For smaller values of V the agreement between the solutions of the asymptotic and original EHL problems is even better. That validates the asymptotic approach on the conceptual level.

Validation of the asymptotic approach on the detailed numerical level involves several steps that include comparison of numerical solutions of the asymptotic and original EHL problems for precritical lubrication regimes and matching some properties of the asymptotic and original EHL problems for both pre- and over-critical lubrication regimes. First, consider the comparison of numerical solutions of asymptotic and original EHL problems for precritical lubrication regimes. To make this comparison it is necessary to go through several steps. First, for the given values of V, Q, and a a solution of the original EHL problem has to be determined. Then the asymptotic equations for $Q_0 = QV^{-1/5}$ and $\alpha_1 = (a + 1)V^{-2/5}$ for the same parameters V, Q, and a have to be solved. After that the comparison of the values of the film thickness H_0 from formula (34) obtained from numerical solutions of the asymptotic and original EHL problems can be done. Similarly, compare the values of exit coordinate c, inlet gap h, and min h obtained from numerical solution of the asymptotic and original EHL problems. Do the comparison for a series of three solutions obtained for $a = -2$, $V = 0.05$ ($\varepsilon_q = V^{2/5} = 0.302$), $Q = 0$ ($Q_0 = 0$), $Q = 1.821$ ($Q_0 = 1$), and $Q = 3.641$ ($Q_0 = 2$) (based on the relationship $Q = Q_0V^{-1/5}$) with the corresponding asymptotic solutions. For all these solutions $\varepsilon_q = V^{2/5} = 0.302$ and $\alpha_1 = -3.314$ (see formula $a = -1 + \alpha_1\varepsilon_q$). The solutions of the original EHL problem are represented in Table 1. In Table 2 the values of $H_{0(asym)}$, $c_{(asym)}$, and min $h_g(s)$ obtained from numerical solution of the asymptotically valid equations in the inlet and exit zones and formulas $H_0 = AV^{3/5}$ and $c = 1 + \beta_1\varepsilon_q$ for the same values of parameters Q, V, and $\alpha_1 = -3.314$ are presented. The step sizes Δx and Δr used for solution of the original and asymptotic EHL problems are chosen in such a way that they provide the same numerical precision (i.e., $\Delta x = \varepsilon_q\Delta r$).

From the data presented in Tables 1 and 2 it is clear that, in spite of the fact that $\varepsilon_q = 0.302$ is not very small, the agreement between the solutions of the original and asymptotically valid equations of the EHL problem is very good. Moreover, the comparison of the values of the film thickness H_0, minimum gap min h, and exit coordinate c obtained from the numerical solution of the original EHL problem and the asymptotic ones shows that the

Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts, Table 1 Parameters H_0, c, and min $h(x)$ obtained from solution of the original EHL problem for cases of $Q = 0$, 1.821, and 3.641

Q (Q_0)	H_0	c	min $h(x)$
0 (0)	0.066	1.030	0.793
1.821 (1)	0.089	1.038	0.771
3.641 (2)	0.110	1.044	0.761

Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts, Table 2 Parameters $H_{0(asym)}$, $c_{(asym)}$, and min $h_g(s)$ obtained from solution of the asymptotic EHL problem for cases of $Q_0 = 0$, 1, and 2

Q_0	$H_{0(asym)}$	$c_{(asym)}$	min $h_g(s)$
0	0.0642	1.0291	0.791
1	0.0883	1.0347	0.770
2	0.1106	1.0391	0.760

Asymptotic Methods for Analyzing Heavily Loaded EHL Contacts, Table 3 Parameters H_0, c, $h(\Delta x/2)$, and min $h(x)$ obtained from solution of the original EHL problem for cases of $V = 0.05$, 0.01 and $Q_0 = 1$, 2

a	V	Q (Q_0)	H_0	c	$h(\Delta x/2)$	min $h(x)$
−2	0.05	1.821 (1)	0.0887	1.0383	24.880	0.7708
−2	0.05	3.641 (2)	0.1097	1.0442	20.268	0.7607
−1.525	0.01	2.512 (1)	0.0337	1.0211	23.473	0.7703
−1.525	0.01	5.024 (2)	0.0419	1.0242	19.023	0.7595

difference is smaller than or equal to 2.7%. The precision of the asymptotic solution becomes even better for smaller values of V.

Now, let's validate the asymptotic approach differently. The asymptotic equations for precritical lubrication regimes (23), (27), and (35) for starved lubrication regimes (or (37) for fully flooded lubrication regimes) and (24), (26), and (36) for the starved lubrication regimes (or (38) for fully flooded lubrication regimes) show that their solutions depend only on two parameters, α_1 and $Q_0 = QV^{1/5}$. Therefore, if the asymptotic analysis for precritical lubrication regimes is valid then it is expected that for different values of parameter V and the same values of parameters α_1 and Q_0 solutions of the original EHL problem are supposed to exhibit a property that the value of coefficient $A = H_0V^{-3/5}$ is constant. Let's examine this statement using two series of solutions of the original EHL problem obtained for $V = 0.05$, $a = -2$ ($\alpha_1 = -3.314$) and $V = 0.01$, $a = -1.525234$ ($\alpha_1 = -3.314$). In both series of calculations it was assumed that $Q_0 = 1$ and $Q_0 = 2$. These values of Q_0 clearly indicate (see the definition of precritical lubrication regimes (42)) that the lubrication regime is precritical. The results of these calculations are presented in Table 3.

From the data of Table 3 for $V = 0.05$, 0.01 and $Q_0 = 1$ it can be determined that $A = H_0V^{-3/5} = 0.5349$ and 0.5336, respectively, while for $V = 0.05$, 0.01 and $Q_0 = 2$ it can be concluded that $A = H_0V^{-3/5} = 0.6617$ and 0.6640, respectively. Therefore, the difference between the corresponding values of constant A is less than 0.34%, which again validates the asymptotic analysis for precritical lubrication regimes.

Finally, examine whether the asymptotic equations derived for precritical regimes can be used for calculations for over-critical lubrication regimes. To do that it is necessary to consider the case of fully flooded over-critical lubrication regime (see the definition of over-critical lubrication regimes (43)) with the viscosity $\mu = \exp(Q_P)$ and the inlet coordinate $a = -1 + \alpha_{10}(VQ)^{1/2}$. Then the same inlet coordinate in a precritical lubrication regime would be $a = -1 + \alpha_1 V^{2/5}$, which means that $\alpha_1 = \alpha_{10}Q_0^{1/2}$, $Q_0 = QV^{1/5}$. As was mentioned earlier, the solution of asymptotic equations for a precritical lubrication regime depends only on the values of parameters α_1 and Q_0. This means that it can be expected to get film thickness $H_0 = AV^{3/5}$, where $A = A(\alpha_{10}, Q_0)$. Therefore, if over-critical lubrications regimes can be described by the asymptotic equations for precritical lubrication regimes, then it is expected that a good match of the values of the lubrication film thickness H_0 obtained from solution of the asymptotic equations for precritical lubrication regimes with the solution of the original EHL problem should be obtained. Consider the case of $a = -2$, $V = 0.05$, and $Q = 10 (Q_0 = 5.4928)$. This is clearly an over-critical lubrication regime because $e^{Q_0} = 242.94 \gg 1$ (see (43)). In this case $\alpha_1 = -3.314$ and from the solution of the asymptotic equations for precritical lubrication regimes follows that $A = 1.0824$ and, therefore, $H_0 = AV^{3/5} = 0.1794$. The solution of the original (not asymptotic) EHL problem for the same values of parameters a, V, Q, gives $H_0 = 0.1743$. The difference between these values of H_0 is 2.9%. Therefore, the asymptotic equations derived for precritical regimes can be used for calculations for over-critical lubrication regimes as well.

Therefore, for heavily loaded pre- and over-critical lubrication regimes (which are determined by the fact that the characteristic size of the inlet zone $\varepsilon_q = V^{2/5} \ll 1$ and $\varepsilon_q = (VQ)^{1/2} \ll 1$, respectively) one can create a map of level curves for $H_0V^{-3/5} = f(Q_0)$, that is, for any values of parameters V and Q for which $Q_0 = QV^{1/5} = const$ there is $H_0V^{-3/5} = const$ (for details see Kudish and Covitch (2010)).

The analysis of the numerical results validates the asymptotic approaches used for the cases of heavily loaded pre- and over-critical lubrication regimes (Kudish and Covitch 2010). Moreover, it reveals that the asymptotic equations obtained for precritical lubrication regimes can be used for calculations of over-critical lubrication regimes.

Key Applications

The presented asymptotic methodology can be successfully applied to studying heavily loaded contacts lubricated by fluids with various non-Newtonian rheologies. Such contacts can be found in bearings, gears, cam/follower systems, and so on. The asymptotic approach provides a distinct and simple way to make generally unstable solutions of the EHL problem for heavily loaded

contacts stable as well as it provides a simple way to determine the suitable size of the numerical step size for given input parameters of the problem (see details in ► Numerical Stability and Precision in Elastohydrodynamic Lubrication (EHL) as well as in Kudish and Covitch (2010)). The asymptotic approach presents simple and asymptotically precise formulas for the lubrication film thickness and, simultaneously, reduces the number of the input parameters affecting the problem solutions in the inlet and exit zones of heavily loaded EHL contacts.

Cross-References

- ► Numerical Stability and Precision in Elastohydrodynamic Lubrication (EHL)
- ► Starvation Effect on Film Thickness in Elastohydrodynamically Lubricated Contacts
- ► Stress-Induced Lubricant Degradation and Viscosity Loss
- ► Thermoelastohydrodynamically Lubricated Contacts with Non-Newtonian Lubricants: Asymptotic Approach

References

J.F. Archard, K.P. Baglin, Elastohydrodynamic lubrication – improvements in analytic solutions. Proc. Inst. Mech. Eng **200**(C4), 281–291 (1986)

E.J. Bissett, D.W. Glander, A highly accurate approach that resolves the pressure spike of elastohydrodynamic lubrication. ASME J. Tribol. **110**(2), 241–246 (1988)

A.W. Crook, The lubrication of rollers II. Film thickness with relation to viscosity and speed. Phil. Trans. Roy. Soc. Lond. **254**(1040), 223–236 (1961)

D. Dowson, G.R. Higginson, *Elastohydrodynamic Lubrication* (Pergamon Press, London, 1966)

A.M. Ertel, Hydrodynamic calculation of lubricated contact for curvilinear surfaces, in *Proceedings of CNIITMASh*, USSR, 1945, pp. 1–64

H.P. Evans, T.G. Hughes, Evaluation of deflection in semi-infinite bodies by a differential method. Proc. Inst. Mech. Eng. **214**(Part C), 563–584 (2000)

A.N. Grubin, The basics of the hydrodynamic lubrication theory for heavily loaded curvilinear surfaces, in *Proceedings of CNIITMASh*, USSR, No. 30, 1949, pp. 126–184

B.J. Hamrock, Pan Ping, Lee Rong-Tsong, Pressure spikes in elastohydrodynamically lubricated conjunctions. ASME J. Tribol. **110**(2), 279–284 (1988)

L.G. Houpert, B.J. Hamrock, Fast approach for calculating film thickness and pressures in elastohydrodynamically lubricated contacts at high loads. ASME J. Tribol. **108**(3), 441–452 (1986)

I.I. Kudish, Hydrodynamic lubrication theory of rolling cylindrical bodies. in *Abstracts of the 2nd All-Union Conference on Elastohydrodynamic Theory of Lubrication and its Practical Application in Technology*, (1976), Kujbyshev, p. 11; in *Proceedings of the 2nd All-Union Conference on Elastohydrodynamic Theory of Lubrication and its Practical Applications in Industry*, Kujbyshev, 1977, pp. 33–38

I.I. Kudish, Elastohydrodynamic problem for a heavily loaded rolling contact. Proc. Acad. Sci. Armen. SSR Mech. **31**(1), 65–78 (1978a)

I.I. Kudish, Asymptotic analysis of a plane non-isothermal elastohydrodynamic problem for a heavily loaded rolling contact. Proc. Acad. Sci. Armen. SSR Mech. **31**(6), 16–35 (1978b)

I.I. Kudish, Asymptotic methods of study for plane problems of the elastohydrodynamic lubrication theory in heavy loaded regimes. Part 1. Isothermal problem. Proc. Acad. Sci. Armen. SSR Mech. **35**(5), 46–64 (1982)

I.I. Kudish, Asymptotic method of study for plane problems of the elastohydrodynamic lubrication theory for heavily loaded regimes. Part 2. Non-isothermal problem. Proc. Acad. Sci. Armen. SSR Mech. **36**(5), 47–59 (1983)

I.I. Kudish, M.J. Covitch, *Modeling and Analytical Methods in Tribology* (Chapman & Hall/CRC Press, Boca Raton, 2010)

M. Van-Dyke, *Perturbation Methods in Fluid Mechanics* (Academic, New York, 1964)

C.H. Venner, A.A. Lubrecht, *Multilevel Methods in Lubrication* (Elsevier, Amsterdam, 2000)

I.I. Vorovich, V.M. Aleksandrov, V.A. Babeshko, *Non-classical Mixed Problems of Elasticity* (Nauka Publishing, Moscow, 1974)

ATF

► Transmission Lubricants

Atmospheric Effects

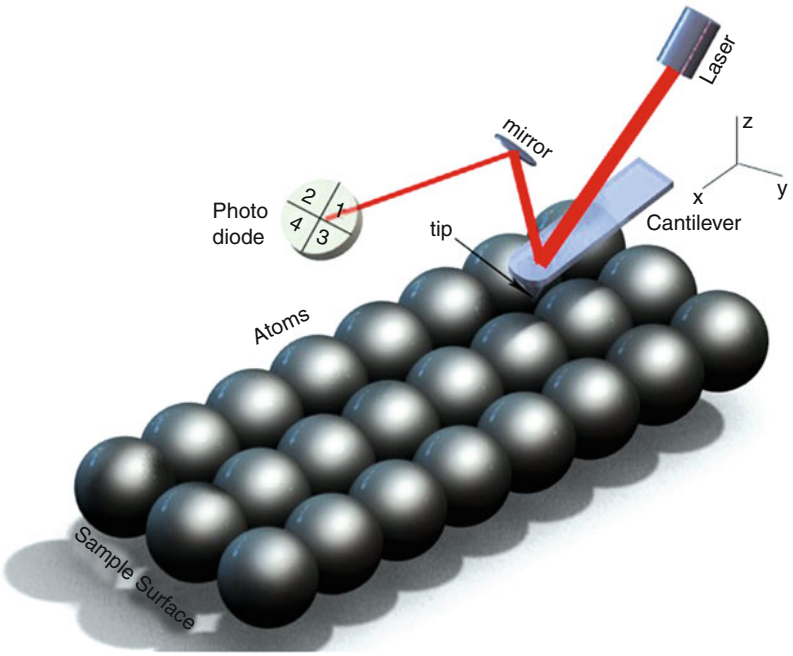
► Vapor Phase Lubrication for Micro-Machines

Atomic Force Microscopy (AFM)

Reza Shahbazian-Yassar
Department of Mechanical Engineering-Engineering Mechanics, Michigan Technological University, Houghton, MI, USA

Definition

The atomic force microscope (AFM) is an instrument for the study of surface properties for both conductive and non-conductive samples (Binnig et al. 1986). AFM is able to build three-dimensional maps of surface properties in both air and liquid environments, with high lateral (<25 nm), vertical (□0.1 A°), and force (□1 *pN*) resolution. Due to such powerful capabilities, AFM is an indispensible technique in tribology to understand surface



Atomic Force Microscopy (AFM), Fig. 1 Schematic of an AFM. The cantilever is deflected by the surface topography of the sample. The signals are detected with a laser-optical set-up. In most AFM systems, the sample rests on a piezotube scanner (not shown in this figure), which allows a scanning motion in *x*- and *y*-directions as well as movement in *z*-direction

topography, adhesion, lubrication, and wear at the atomic scale (Adams et al. 2001; Bhushan et al. 1995).

Scientific Fundamentals

Mechanism

AFM operates by the movement of a cantilever that deflects when interacting with the sample surface. The cantilever scans the surface by means of a piezoactuator. The cantilever deflection is measured to reproduce the sample topography. The most common approach to detect cantilever deflections is the optical lever method that focuses a laser beam on the back side of the cantilever and detects the reflected beam by a position sensor, which is usually a quartered photodiode (Fig. 1). When the cantilever moves up and down on the surface, the reflected laser beam projects on the photodiode correspondingly on different locations. The interaction between laser and photodiode results in electrical potential, which can be used to generate useful information about the sample surface.

Operational Modes

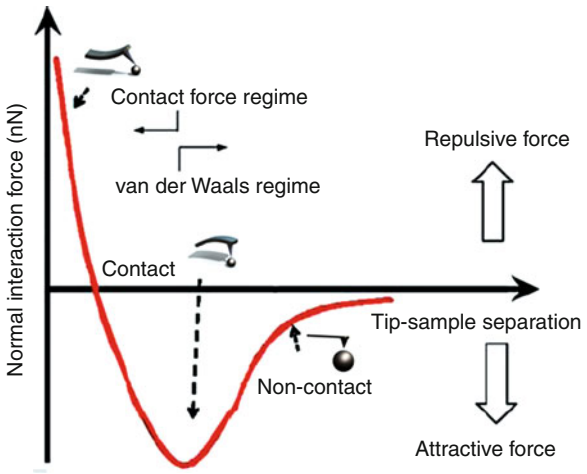
AFM typically operates in three distinct modes: (a) contact mode, (b) non-contact mode, and (c) tapping mode. In contact mode, the cantilever tip is in continuous contact with the sample surface. In non-contact mode, the cantilever tip oscillates about 50–150 Å above the sample surface to detect surface forces acting between the tip and the sample. In tapping mode, the tip intermittently contacts the surface and oscillates typically at amplitudes of 100–200 nm. The frequencies of oscillation are 50–500 kHz in air and approximately 10 kHz in fluids.

The interaction forces between the tip and sample in all three modes can be distinctly identified on a force–displacement curve as shown in Fig. 2. In short distances above sample surface, the van der Waals forces dominate. By decreasing this distance, the attractive force increases until AFM tip contacts the surface. At this position, the attractive force is at the maximum. As the distance becomes even smaller, the nature of this force transforms to repulsive due to the interaction between electron clouds of AFM tip and those of sample surface.

Key Applications

Surface Roughness

Determination of surface roughness by AFM is important to the study of surface topography at micro- and



Atomic Force Microscopy (AFM), Fig. 2 Interaction force versus distance between AFM tip and sample surface

nano-scale. For roughness measurement, the detected variation of cantilever deflection in the vertical position (the *z* axis) reflects the topography of the surface. AFM can be used in all three modes: contact, tapping, and noncontact modes to obtain the surface topography maps. In contact and tapping modes, the cantilever is in contact with the surface (continuously or intermittently). Figure 3 shows the three-dimensional surface profiles of a titanium sample in contact and tapping modes. Due to strong repulsive forces (10^{-7} N), an external force needs to be exerted on the cantilever to maintain contact between the cantilever and surface. This may result in the change of surface topography for soft surfaces such as polymer films. In such cases, the roughness of scanned surfaces can be slightly higher in tapping or contact mode versus the non-contact mode.

There are several parameters with respect to the choice of scanning parameters and AFM tip shape (Kiely and Bonnell 1997) that need to be considered for the roughness calculations. The typical cantilever probes used in AFM imaging are not ideally sharp. As a consequence, an AFM image does not reflect the true sample topography, but rather represents the interaction of the probe with the sample surface. This is called tip convolution. In addition, due to the limitation of AFM piezostage movement in X and Y directions, the calculated roughness can vary depending on the size of the scanned area. For instance, an increase in surface roughness values by increasing the size of scan area has been reported (Boussu et al. 2005). Therefore, it is critical to carefully select scanning parameters and take tip convolution into consideration.

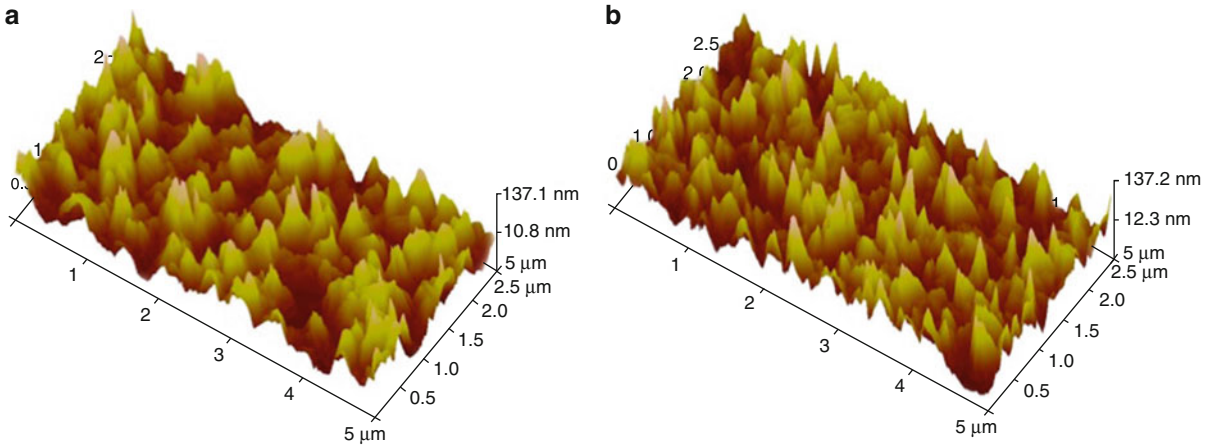
Surface Adhesion

In recent years, AFM has become a powerful tool in the study of surface adhesion at nano-scale (Burns et al. 1999; Brogly et al. 2009; Jones et al. 2002). Adhesion forces are determined by measuring the attractive portion of the interaction force versus distance. As the AFM tip is pulled out of the contact from a surface, adhesion is calculated as a negative (attractive) force, which occurs just before the tip comes free from the surface. The forces involved include van der Waals forces, hydrogen bonds, electrostatic forces, and chemical bonding in some cases. In order to carry out a quantitative analysis, various experimental details should be taken into consideration, such as the selection of tips that possess apex spherical shape, cantilever spring constant, tip radius, linearity of photodiodes, cantilever and piezo stabilities, and tip contamination (Noel et al. 2004).

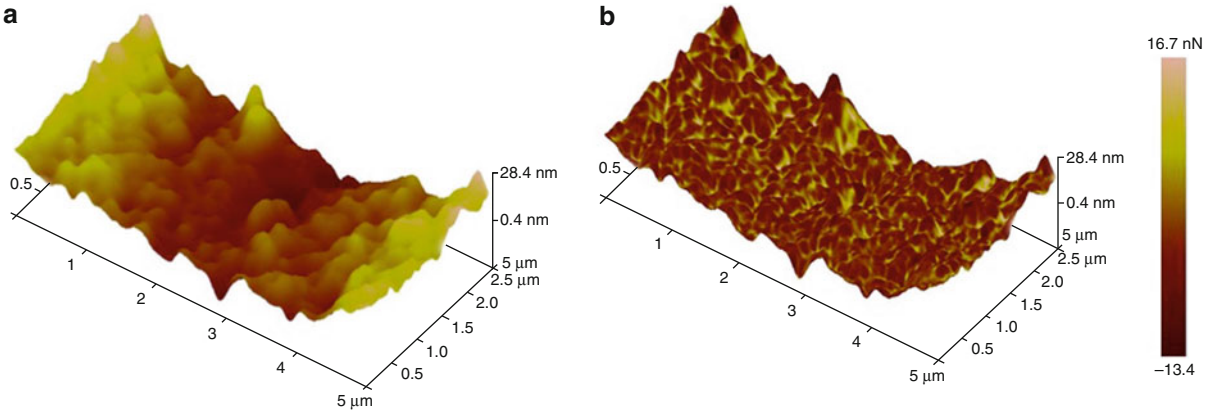
The tapping mode has typically been used to measure adhesion properties. This is due to the facts that (1) when the tip contacts the surface, it has sufficient oscillation to overcome the tip-sample interaction; and (2) the AFM cantilever is not pulled sideways by shear forces since the applied force is always vertical. However, one should note that strong adhesion forces may influence the measurement of surface topography. In the presence of strong adhesion forces, other AFM modes should also be used to map the topography in order to deconvolute the effect of adhesion on topography measurements. Polymeric films are favorable materials to study adhesive properties with AFM. Figure 4 shows the 3-D height profile and adhesion map of a polymer thin film. The adhesion map reveals that there is variation of surface adhesion from −13 to 69 nN on this particular surface.

Surface Wear

As the dimensions of component and loads continue to decrease, scratch/wear resistance measurement at the micro- and nano-scale becomes increasingly important (Bhushan and Ruan 1994). The wear coefficient, defined by the Archard's wear equation (Archard 1953) provides a somewhat accurate measurement to gauge the wear resistance. For scratch and wear resistance studies, the



Atomic Force Microscopy (AFM), Fig. 3 Three-dimensional surface height profiles of a titanium sample in (**a**) contact and (**b**) tapping modes

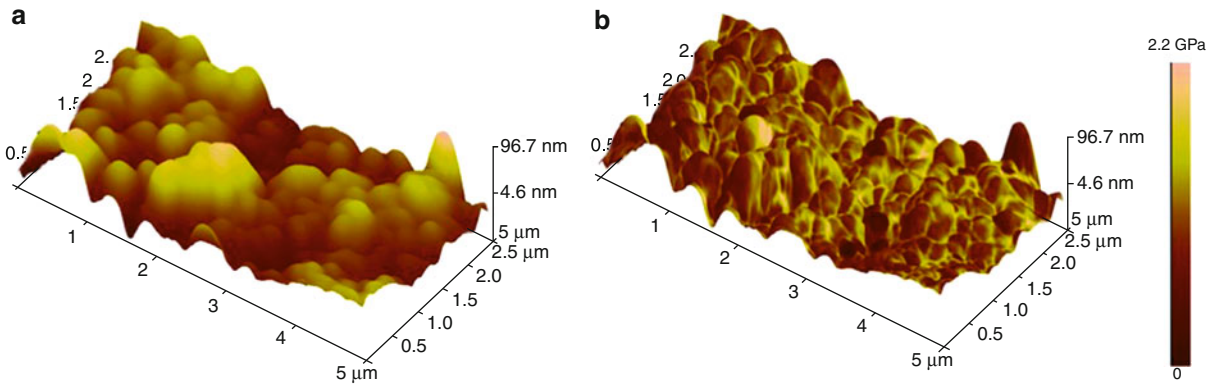


Atomic Force Microscopy (AFM), Fig. 4 (**a**) Three-dimensional surface height profile and (**b**) adhesion maps of a polymer thin film

sample is generally scanned using an AFM equipped with a diamond tip mounted on the stiff cantilever (Gnecco et al. 2002; Yasui et al. 2009; Kato et al. 2003). Scratch experiments are generally performed at ramped loads to obtain the friction signal during the scratching operation and to determine a critical load at which failure initiates.

Surface Mechanics

Due to its very fine force sensing capabilities and ultra small tip radius, AFM has also been used as a powerful tool to characterize mechanical properties of various surfaces (Burnham and Colton 1989). In this application, the vertical position of the piezo (Z_p) and the deflection of cantilever (Z_c) are recorded as the tip approaches and retracts the sample. When a normal force (F) is applied, the cantilever bends and the position of reflected beam changes on the detector. The deflection of a cantilever (E = elastic modulus) with rectangular cross section of thickness (t), width (w), and length (L) is given by $Z_c = \frac{4FL^3}{Ewt^3}$. Knowing the spring constant of the cantilever (k_c), the force is calculated by $F = k_c Z_c$. Figure 5 shows a three-dimensional height profile and elastic modulus map of sintered carbon particles, indicating variation of elastic modulus ranging from 0.5 to 2 GPa.



Atomic Force Microscopy (AFM), Fig. 5 (**a**) Three-dimensional surface height profile and (**b**) three-dimensional elastic modulus maps of sintered carbon particles

Cross-References

► Surface Analysis Using Contact Mode AFM
► Surface Analysis Using Dynamic AFM
► Surface Characterization and Description

References

J. Adams, L. Hector, D. Siegel, H. Yuand, J. Zhong, Adhesion, lubrication and wear on the atomic scale. Surf. Interface Anal. **31**, 619–626 (2001)

J. Archard, Contact and rubbing of flat surfaces. J. Appl. Phys. **24**, 981 (1953)

B. Bhushan, J. Ruan, Atomic-scale friction measurements using friction force microscopy. Part II. Application to magnetic media. ASME J. Tribol. **116**, 389–396 (1994)

B. Bhushan, J.N. Israelachvili, U. Landman, Nanotribology: friction, wear and lubrication at the atomic scale. Nature **374**, 607–616 (1995)

G. Binnig, C.F. Quate, C.H. Gerber, Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)

K. Boussu, B. Van der Bruggen, A. Volodin, J. Snauwaert, C. Van Haesendonck, C. Vandecasteele, Roughness and hydrophobicity studies of nanofiltration membranes using different modes of AFM. J. Colloid Interface Sci. **285**, 632–638 (2005)

M. Brogly, H. Awada, O. Noel, Contact atomic force microscopy: a powerful tool in adhesion science, in *Nanoscience and Technology*. Applied Scanning Probe Methods, vol. XI (Springer, Berlin, 2009), p. 73

N.A. Burnham, R.J. Colton, Measuring the nanomechanical properties and surface forces of materials using an atomic force microscope. J. Vac. Sci. Technol. **7**, 2906 (1989)

A. Burns, J. Houston, W. Carpick, A. Michalske, Molecular level friction as revealed with a novel scanning probe. Langmuir **15**, 2922 (1999)

E. Gnecco, R. Bennewitz, E. Meyer, Abrasive wear on the atomic scale. Phys. Rev. Lett. **88**, 215 (2002)

R. Jones, H. Pollock, J. Cleaver, C. Hodges, Adhesion force between glass and silicon surfaces in air studied by AFM: effect of relative humidity, particle size, roughness and surface treatment. Langmuir **18**, 8045 (2002)

Z. Kato, M. Sakairi, H. Takahashi, Nanopatterning on aluminum surfaces with AFM probe. Surf. Coat. Technol. **169**, 195 (2003)

J. Kiely, A. Bonnell, J. Vac. Sci. Technol. **15**, 1483 (1997)

O. Noel, M. Brogly, G. Castelein, J. Schultz, In situ determination of the thermodynamic surface properties of chemically modified surfaces on a local scale: an attempt with atomic force microscope. Langmuir **20**, 2707 (2004)

N. Yasui, H. Inaba, K. Furusawa, M. Saito, N. Ohtake, Characterization of head overcoat for 1 Tb/in2 magnetic recording. IEEE Trans. Magnet. **45**, 805 (2009)

Atomic Lattice Stick-Slip

► Atomic-Level Stick-Slip

Atomic Layer Deposition (ALD)

Cesar Augusto Duarte Rodriguez,
Germano Tremiliosi-Filho
Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, São Paulo, Brazil

Synonyms

ALE - atomic layer epitaxy; ALP – atomic layer processing

Definition

ALD is a coating method that has the capability to control the thickness, in atomic scale, of ultrathin uniform films formed by sequential self-limiting surface reaction of adsorbed gas layers.

Scientific Fundamentals

The discovery of new materials with good surface properties has had a dramatic impact on technological progress. Depending on the type of film material and its applications, various deposition techniques may be employed. Specifically, ALD refers to the method whereby film growth occurs by surface reaction after exposing the surface substrate to the gas reactants individually in a sequential manner. This technique provides chemical modification of the surface by film deposition with atomic precision. The advantages of ALD in relation to other deposition techniques are:

- Inherent control of the film thickness
- Film thickness is determined by number of sequential surface reactions
- Unique self-limiting growth process
- High conformity and large area uniformity
- Porous structure or particle beds can be uniformly coated by an ultra-thin layer obtained by ALD technique

A selection of materials can be deposited by ALD, including selenides, sulfides, nitrides, and oxides. The main applications are in the semiconductor industry, sensors, solar panels, magnetic heads, memory, fuel cells, flat panel displays, primer layers, protective layers, optical filters, biomedical coatings, UV blocking layers, electroluminescent displays, corrosion prevention, and others. The emerging applications, including surface passivation, wear resistance, and solid lubricant layers, produce good corrosion, mechanical, and nano-tribological properties. The ALD technology was developed in the 1970s for production of electroluminescent flat-panel displays (Suntola and Antson 1977; Suntola and Hyvarinen 1985). This was the first industrial use of ALD and its production continues today (Niinisto et al. 2009).

Tribological Applications

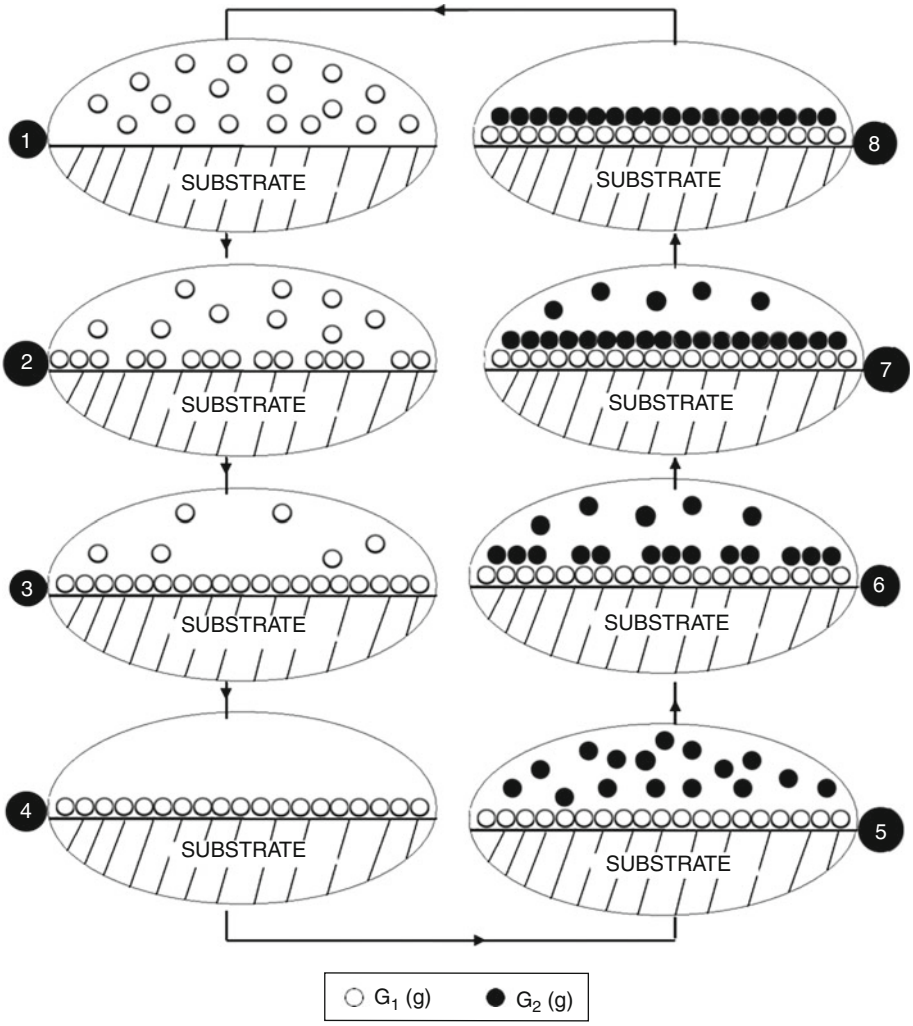
One basic condition for successful use of the ALD method, especially considering tribological applications, is that the binding energy of the first film monolayer deposited on the surface is higher than the binding energy of subsequent layers on top of the first. The greater the difference between the bond energy of the first monolayer and the bonding energy of the subsequent layers, the better the tribological properties.

Typical film characteristics for tribological application, and specifically for the ALD process, are atomic level control of film composition and thickness, stepped surface, and uniformly covered surface defects, including large-area surfaces. Because of the self-limiting adsorption principle of ALD and the benefits it brings, ALD is rather insensitive to the surface topology and is suitable for depositing conformal films on nano-size objects. Thus, the ALD coatings have attracted attention in tribological applications due to properties such as high hardness, low friction coefficient, high wear resistance, and chemical inertness combined in a unique ultrathin surface layer. The potential application of ALD in tribology has been recognized for some time, especially for ceramics films such as Al_2O_3 and ZrO_2 deposited on macroscopic or nano-particles surfaces. Incorporation of nitrogen in metallic (e.g., TiN) or amorphous (e.g., CN_x) films also improves their properties in terms of wear resistance.

The quantitative and qualitative wear characteristics of different ALD films can be examined using an atomic force microscope (AFM) (Bhushan et al. 1995; Bhushan 2005).

Deposition Process

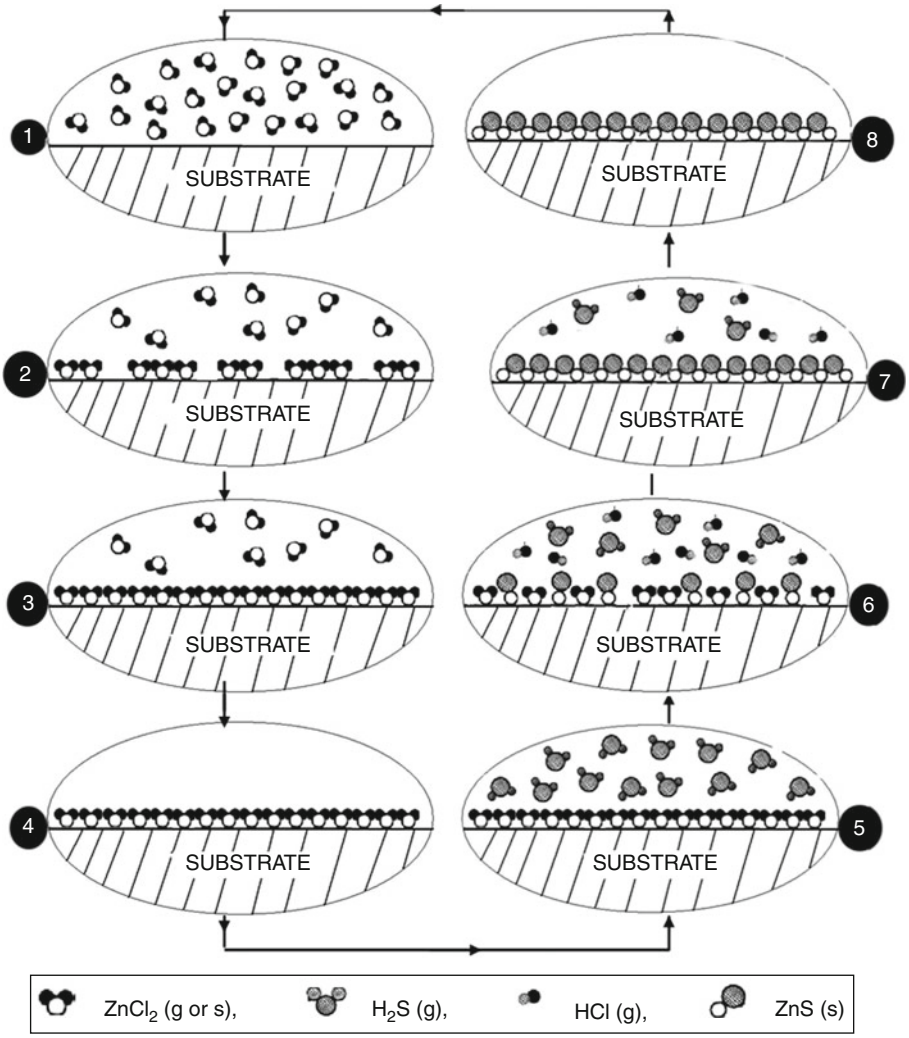
The ALD method is composed of sequential steps involving a deposition of a film, one layer at a time, by surface-controlled growth using surface reactions between two adsorbed layers of gases called precursors. The main steps are indicated as follows: (1) the bare substrate surface is exposed to the first gaseous precursor molecule, $G_1(g)$, (elemental vapor or volatile compound) in excess and with the temperature and gas flow adjusted so the reactant is chemisorbed onto the surface without thermal decomposition; (2) some time is spent for surface saturation by the formation of a precisely dosed compact surface monolayer of adsorbed gas, $G_1(s)$; (3) The excess reactant, $G_1(g)$, which is in the gas phase or physisorbed on the surface, is then purged out of the chamber, leaving behind the chemisorbed monolayer of $G_1(s)$; (4) the other reactant gas, $G_2(g)$, is then introduced into the reaction chamber to chemisorb onto the previously deposited monolayer of $G_1(s)$ and undergoes a surface reaction with this first adsorbed reactant, $G_1(s)$; (5) surface saturation by formation of a compact surface monolayer of a product formed by the surface reaction, $G_1(s) + G_2(s) \rightarrow G_1(s)–G_2(s)$, once all of the surface sites are reacted, no more reaction takes place, thus, the reaction is self-limited, and this results in the formation of a molecular solid monolayer film ($G_1(s)–G_2(s)$); (6) the reaction chamber is purged by pumping away the excess of $G_2(g)$ and the reaction by-products to prevent unwanted gas-phase reaction; finally, (7) the process can be repeated sequentially until the desired film thickness is achieved. Because one step cycle is constituted by the reaction of each pair of the adsorbed monolayer of gases introduced into the reaction chamber that produces exactly one monolayer of the resulting film, the thickness



Atomic Layer Deposition (ALD), Fig. 1 (*1*) Introduction of the first precursor gas $G_1(g)$ into the reaction chamber, (*2*) Chemisorption of $G_1(g)$ onto the bare substrate surface, (*3*) Surface saturation by adsorption limited to one monolayer of $G_1(s)$, (*4*) Purging of the reaction chamber to remove the non-absorbed precursor $G_1(g)$ and others possible reaction by-products, (*5*) Instruction of the second precursor gas $G_2(g)$ into the reaction chamber, (*6*) Chemisorption of precursor $G_2(g)$ onto the $G_1(s)$ surface, (*7*) Surface saturation by adsorption limited to one monolayer of $G_2(s)$ on $G_1(s)$ monolayer previously adsorbed and surface reaction in order to form the molecular monolayer film $G_1(s)$–$G_2(s)$ and (*8*) Purging of the reaction chamber to remove the non-absorbed precursor $G_2(g)$ and other possible reaction by-products

of this film may be precisely controlled by the number of deposition cycles. The cyclic repetition of steps (1) to (6) enables the formation of a thin film constructed of as many molecular packed layers as wanted. This process is depicted in more detail in Fig. 1 and can be cyclically repeated as many times as necessary to form the desired final film structure.

In general, the adsorbed species, $G_1(s)$ and $G_2(s)$, react at the surface assisted by certain activation energy (temperature) without interference of a gas-phase reaction. Thus, the ALD offers a controlled surface structure with accurate control and reproduction of the film thickness at atomic level, independent of substrate geometry, only by counting the number of deposition cycles. There is no limitation for coating ultra-fine nano-substrate particles or porous substrate. It is important to recognize that the inherent control of the ALD method allows an effective film growth over particle surface or porous matrix. No other known process allows such conformal film growth on porous substrate or individual particle, even in



Atomic Layer Deposition (ALD), Fig. 2 (*1*) Introduction of precursor $ZnCl_2$ (g) into the reaction chamber, (*2*) Chemisorption of $ZnCl_2$ (g) onto the bare substrate surface, (*3*) Surface saturation by adsorption limited to one monolayer of $ZnCl_2$ (s), (*4*) Purging of the reaction chamber to remove the non-absorbed precursor $ZnCl_2$ (g), (*5*) Introduction of precursor H_2S (g) into the reaction chamber, (*6*) Chemisorption of H_2S (g) onto the surface monolayer previously adsorbed $ZnCl_2$ (s) and/or reaction between H_2S (g) and $ZnCl_2$ (s), (H_2S (g) + $ZnCl_2$ (s) → ZnS (s) + 2 HCl (g)), (*7*) Surface saturation by ZnS (s) limited to one monolayer, and, (*8*) Purging of the reaction chamber to remove the non-absorbed precursor H_2S (g) and the reaction by-product (HCl(g))

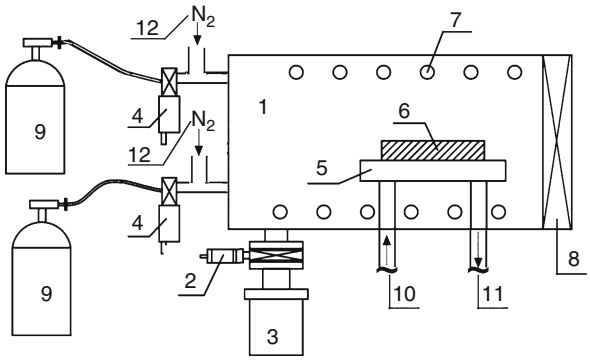
presence of particle aggregates where no sintering between particles is observed. Each reaction cycle adds a molecular layer to the film up to 3 Å in thickness; it may take a few seconds to grow each molecular layer on flat substrate or a few minutes to saturate the internal porous walls of a porous substrate. ALD processing ensures that a non-granular, pinhole-free, uniform, and conformal coating of controlled thickness is formed. In general, the ALD uses a surface self-terminating reaction between two adsorbed chemicals (G_1(s) + G_2(s)), and because of this, only one monolayer of film is formed and remains on the surface after each exposure to the reactant gases. Repeating this cycle leads to a controlled layer-by-layer growth. ALD is a surface-controlled process where parameters such as temperature, chemical nature of the precursors, and substrate may have marked influence. Various volatile compounds can be used as metal source materials in ALD, such as halides ($ZrCl_4$), alkoxides ($Zr(O(CH_3)_3)_4$), organometallics ($Al(CH_3)_3$), and amido complexes ($Zr(N(CH_3)_2)_4$). Figure 2 presents the formation of ZnS (s) starting from

$ZnCl_2$ (g) and H_2S (g). ALD can be used to deposit several kinds of thin films, including oxides (Al_2O_3, TiO_2, ZnO, SiO_2), nitrides (BN, TiN, NbN, WN, Si_3N_4), sulfides, metals (Pt, Ir, Pd), doped materials (ZnS:Mn, ZnO:Al), silicates, polymers (polyimides), and composite films (ZnO/Al_2O_3).

Apparatus

According to the working pressure, the method of adding the precursor gases, = or the sources of reactants, several equipment types are proposed. However, the typical ALD apparatus requires a single air-isolated chamber. Before starting the sequential ALD process, the sample substrate surface must be thermally treated in order to clean and reach a minimum surface energy according to the stringent requirements of surface science for impurity control before proceeding to the deposition process. Air contact with the sample surface is strictly avoided during the thermal treatment. The sample heating treatment is only done under vacuum or in an inert or controlled gas atmosphere. A simplified schematic view of an ALD reaction chamber is shown in Fig. 3.

Because the system must be evacuated during the sequential cyclic deposition procedure, a vacuum pump (3) is used. This pump is connected to the reaction chamber (1) by an appropriated throttle valve (2). The vacuum pump is used to control the system pressure and gas flow and to ensure purging of the chamber after each deposition cycle. The reaction chamber contains precise temperature controls with heaters (7) and a cooled sample support (5). The sample support cooling is achieved by passing cold water (10, 11). After the introduction and accommodation of the sample (6) on the sample support, the reaction chamber door (8) is closed and the system is evacuated to proceed the thermal treatment of the sample. The inlet dosing gas valve (4) fills the reaction chamber with the precursor gas or volatile compound up to the desired pressure at a given adsorption temperature. At least two inlet dosing valves (4) are attached to the reaction chamber and connected to the gas or volatile compound vessels (9). The vessels can be heated to a desired temperature. The first precursor gas or volatile compound is injected into the reaction chamber and the sequential steps illustrated in Fig. 1 are carried out to perform the film deposition process. Representative operational conditions are pressure of 0.1–5 mbar, temperature of 60–500°C, and gas flow of 0.3–1 SLM (standard liters per minute). Besides heating the sample to promote the surface chemical reaction, other forms of activation energy can be used, including plasma, UV, and



Atomic Layer Deposition (ALD), Fig. 3 (*1*) Main chamber, (*2*) throttle valve, (*3*) vacuum pump, (*4*) gas inlet valve, (*5*) sample support, (*6*) sample, (*7*) heating elements, (*8*) main chamber door, (*9*) gas or volatile compound vessel, (*10*) and (*11*) cold water in and out, and, (*12*) nitrogen supply

visible light. Figure 4 illustrates the ALD process for deposition of TiN on a Si substrate using a plasma-assisted step to split H_2 and N_2 molecules into single element radicals.

An important aspect of ALD is to prevent the surface from being exposed to both reactant gases simultaneously. Between exposures, the chamber should be purged by pumping it out with the vacuum pump or an inert gas (N_2 (12)) to sweep away any remaining reactant gas from the system. Simultaneous exposure to reactant gases would result in undesirable chemical vapor deposition (CVD) without control over the surface reaction and film growth.

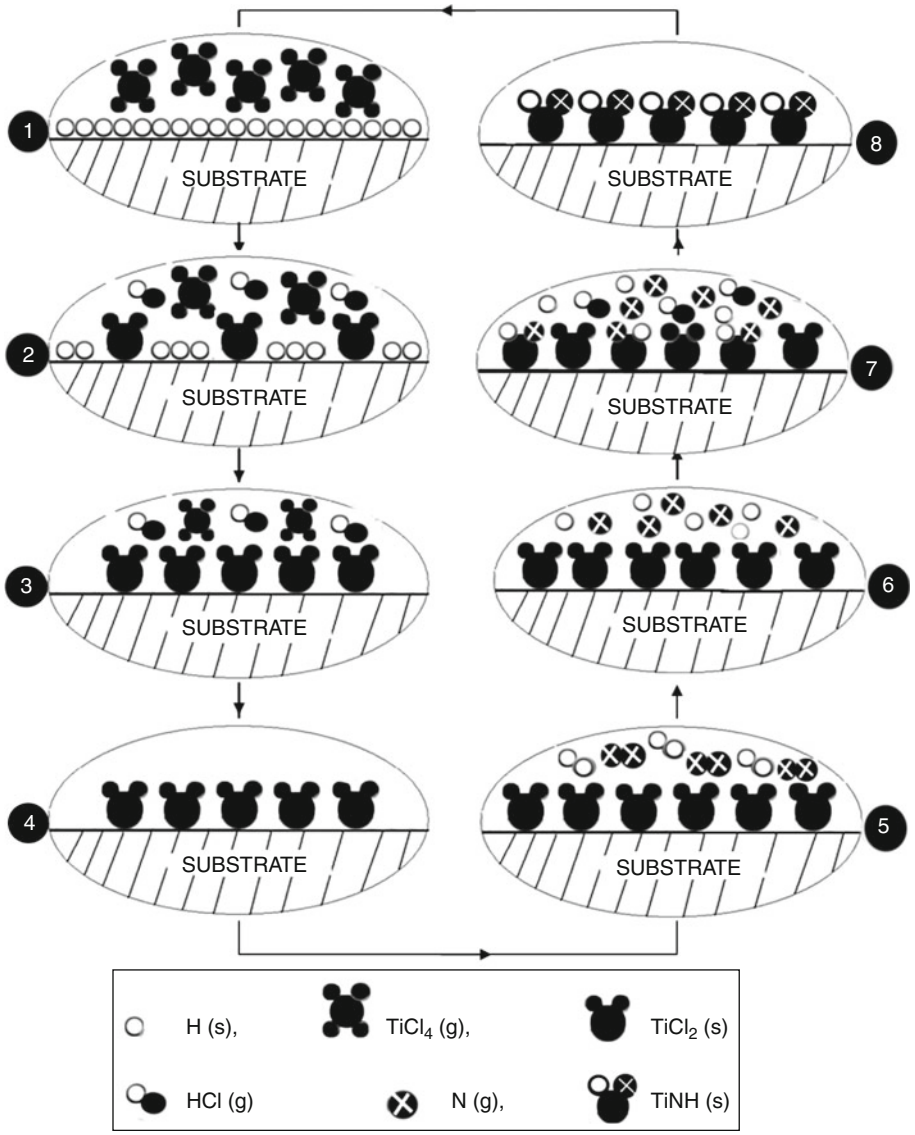
Requirements for ALD Film Growth

Successful ALD growth depends on several parameters:

- The physical and chemical properties of the precursor gases
- The form of introducing the reactant gases into the reactor
- The interaction of the reactant gases with the substrate and with each other
- Volatility of the adsorbed monolayer itself

Volatile precursor compounds used in ALD should be either an inorganic, organic, or organometallic compound. It must be a gas or must volatilize at a temperature that produces high vapor pressure, sufficient to cover the substrate surface to chemisorb a monolayer in a reasonable time interval.

It is undesirable that the chemical reaction between the precursor gases occurs prior to chemisorption of one of the gases onto the substrate surface. This can be avoided



Atomic Layer Deposition (ALD), Fig. 4 (*1*) Introduction of precursor $TiCl_4$ (g) into the reaction chamber, (*2*) Chemisorption and reaction of $TiCl_4$ (g) precursor with H_{ads} (s) from the functional surface group –OH, producing HCl ($TiCl_4$ (g) + 2 H_{ads} (s) → $TiCl_2$ (s) + 2 HCl (g)), (*3*) Surface saturation by $TiCl_2$ (s) limited to one monolayer, (*4*) Purging of the reaction chamber to remove the non-absorbed precursor $TiCl_4$ (g) and the reaction by-product HCl (g), (*5*) Introduction of precursors H_2 (g) and N_2 (g) into the reaction chamber, (*6*) Plasma switched on for the H_2 (g) and N_2 (g) gas molecules split into single element radicals ($H^{\bullet}$ (g) and $N^{\bullet}$ (g)), (*7*) $H^{\bullet}$ (g) reacts with the Cl from the $TiCl_2$ (s) surface monolayer previously deposited, forming HCl (g) and remain in the surface TiCl (s), additionally, the radicals $H^{\bullet}$ (g) and $N^{\bullet}$ (g) react with the TiCl (s) forming TiNH (s), and surface saturation by TiNH (s) limited to one monolayer, and (*8*) Purging of the reaction chamber to remove all radicals and the reaction by-product (HCl(g)), leaving the TiNH (s) with excess of adsorbed H, in order to repeat the process

by purging the reaction chamber before adding the reactants. The purge procedure also cleans the system by removing the excess of non-reacted molecules or by-products.

The self-controlled film growth layer-by-layer and the film quality are achieved only if the chemisorptions occur before a possible decomposition of the precursors on the substrate.

In the ALD process, temperature is one of the most important parameters. If the temperature is lower than necessary, the required activation energy for the surface reaction may not be attained. If the temperature is higher than desired, the reactant may decompose. Only in the appropriate temperature range is the adsorbed monolayer formed and the film quality achieved. In general, the ideal film growth occurs when the temperature is adjusted for a complete saturation of the surface.

The number of binary compounds, most notably oxides, deposited by ALD is constantly increasing for tribological applications. On the other hand, progress with ternary compounds is not observed in the literature and it is questionable whether ALD can effectively compete in this area with other techniques such as CVD.

Key Applications

One of the most important tribological applications of the ALD method is in the friction and wear of coated microelectromechanical system (MEMS) surfaces. Developments in the field of micro-technology have spurred extensive research to understand wear at the microscopic scale. As a result of these efforts, applications of protective coatings in precision systems such as hard disk drives, head/disk interfaces, and micro-rotor spinning have adequately overcome the reliability problems of MEMS devices. To achieve low friction and sufficient wear resistance in MEMS devices, hard materials such as W, metal oxides (Al_2O_3, ZrO_2, and TiO_2) grown by ALD have been used as protective coatings.

Microelectromechanical Systems

MEMS comprise miniaturize mechanical devices produced on length scales from fractions of a micrometer to millimeters. These devices are applied in a wide variety of technologies, including optical switching, mechanical actuators, and others. Due to the small size of these devices, surface-to-volume ratio increases; thus, the inertial forces become insignificant compared with surface forces. Considering that in many applications the MEMS devices with a silicon surface present harsh contacting interfaces, adhesion, friction, and wear dominate their performance and reliability. To relieve these problems, the critical surfaces of MEMS devices can be coated in order to reduce friction and wear and to prevent adhesion by depositing a hard layer as Al_2O_3 or a solid lubricating compound such as WS_2 or MoS_2.

The three-dimensional structure of MEMS makes coatings difficult. The ALD method must deposit a nano-film of equal thickness, composition, and properties on the surface of these devices. Most other processes, e.g., evaporation, sputtering, and CVD, are not able to produce good quality, homogeneous films.

Thin coatings of hard materials such Al_2O_3, ZrO_2, or SiC should protect moving parts of MEMS devices, increasing the lifetime. Thus, the application of an Al_2O_3 (alumina) film by ALD of nominal 10–80 nm thickness on the surface of a MEMS device is potentially useful as a wear-resistant coating. The reaction sequence for deposition of Al_2O_3 on silicon involving trimethyl aluminum and water as precursor is as follows

a formation of AlOH (s) surface functional species:

$$Al(CH_3)_3\ (g) + S\text{–}OH\ (s) \rightarrow S\text{–}O\text{–}Al(CH_3)_2\ (s) + CH_4\ (g)$$

$$2\ H_2O\ (g) + S\text{–}O\text{–}Al(CH_3)_2\ (s) \rightarrow S\text{–}O\text{–}Al(OH)_2\ (s) + 2\ CH_4\ (g)$$

$$Al(CH_3)_3\ (g) + 2\ H_2O\ (g) + S\text{–}OH\ (s) \rightarrow S\text{–}O\text{–}Al(OH)_2\ (s) + 3\ CH_4\ (g)$$

b film growth:

$$2\ S\text{–}O\text{–}AlOH\ (s) + 2\ Al(CH_3)_3\ (g) \rightarrow 2\ S\text{–}O\text{–}[AlOAl(CH_3)_2]\ (s) + 2\ CH_4\ (g)$$

$$2\ S\text{–}O\text{–}[AlOAl(CH_3)_2]\ (s) + 3\ H_2O\ (g) \rightarrow Al_2O_3\ (s) + 2\ S\text{–}O\text{–}AlOH\ (s) + 4\ CH_4\ (g)$$

$$2\ Al(CH_3)_3\ (g) + 3\ H_2O\ (g) \rightarrow Al_2O_3\ (s) + 6\ CH_4\ (g)$$

where surface species are denoted by S–O. Each reaction involves a gas-phase precursor and a surface functional group. The initial surface functionality is formed by the adsorption of water vapor on the silicon surface forming a hydroxyl adsorbed group (S–OH). After sequential exposures of trimethyl aluminum and H_2O, the thickness of the film is ca. 0.12 nm per coating cycle. Nominally 10–80 nm of alumina was deposited using 100–800 cycles of trimethyl aluminum/water exposure.

Nano-Particles

The ability to coat fine particles with thin layers of other material, avoiding particle agglomeration, is one of the advantages of the ALD process. One can envision many applications in this area, such as the nano-coating of individual polymer particles with various ceramic materials to improve impact strength or other structural properties. Nano-particles can also be encapsulated in an ultra-thin layer to acquire the desirable properties to interact with the surrounding environment. Finally, ALD can be directly scalable from the laboratory reactor to a pilot plant for processing batches of hundreds of grams of nano-coated particles.

Conclusion

Friction and wear of moving components in mechanical systems is a critical problem that influences performance and life. ALD is suitable for deposition of friction- and wear-reducing films on surfaces and adds extra extension to the ability to tailor the properties of surfaces for

tribological applications. With advances in the scaling down of devices, and due to its benefits over other conventional deposition techniques, ALD has been proposed as one of the most promising deposition techniques to enable nanoscale device fabrication.

Cross-References

- ► Atomic Layer Deposition (ALD)
- ► Bonding at Surfaces/Interfaces
- ► Chemical Vapor Deposition Processes for Boundary Lubricants
- ► Multiplex Coatings
- ► Nanocomposite Coatings
- ► Self-assembled Monolayers
- ► Surface Nanocrystallization and Hardening (SNH)

References

B. Bhushan, Micro/nanotribology and materials characterization studies using scanning probe microscopy, in *Nanotribology and Nanomechanics. An Introduction*, ed. by B. Bhushan (Springer, Berlin, 2005), pp. 315–387. Cap. 8

B. Bhushan, J.N. Israelachvili, U. Landman, Nature **374**, 607 (1995)

J. Niinisto, K. Kukli, M. Heikkila, M. Ritala, M. Leskela, Adv. Eng. Mater. **11**, 223 (2009)

T. Suntola, J. Antson, *Method for Producing Compound Thin Films*, U.S. Patent 4, 058,430, 1977

T. Suntola, J. Hyvarinen, Annu. Rev. Mater. Sci. **15**, 177 (1985)

Atomic Stick-Slip

► Atomic-Level Stick-Slip

Atomic-Level Stick-Slip

Robert W. Carpick[1], Ashlie Martini[2], Rachel J. Cannara[3]
[1]Mechanical Engineering & Applied Mechanics, University of Pennsylvania, Philadelphia, PA, USA
[2]School of Engineering, University of California Merced, Merced, CA, USA
[3]Center for Nanoscale Science and Technology, National Institute of Standards and Technology, Gaithersburg, MD, USA

Synonyms

Atomic lattice stick-slip; Atomic stick-slip; Atomic-scale stick-slip

Definition

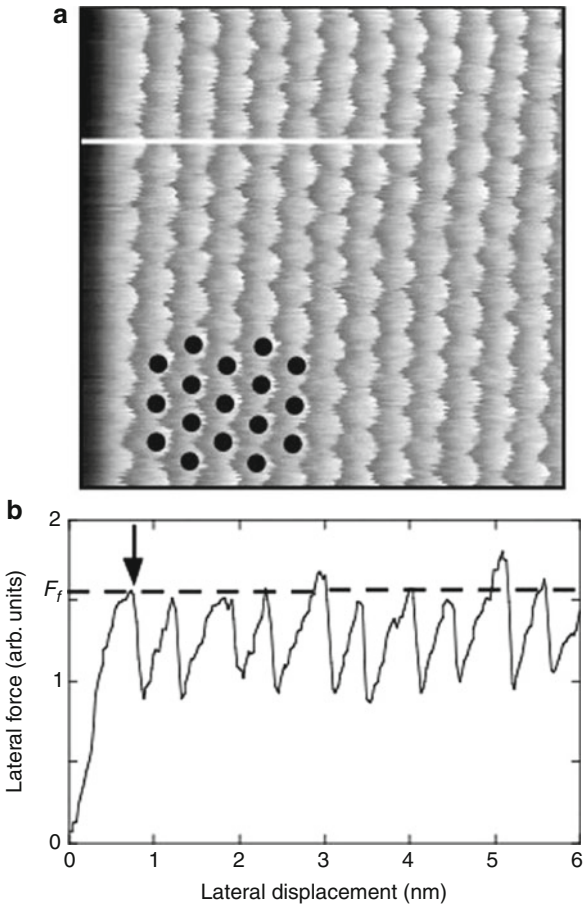
Atomic-level stick-slip refers to the behavior of a sliding interface, usually an atomic force microscope tip sliding along a crystalline surface, whereby the tip sticks and then slips laterally with respect to the surface in a periodic fashion. The periodicity coincides with the surface lattice.

Scientific Fundamentals

History of Atomic-Level Stick–Slip

Atomic-level stick-slip friction is a widely observed phenomenon in atomic force microscopy (AFM) and has been reviewed in detail in the literature (Morita et al. 1996; Szlufarska et al. 2008; Mate 2008). Atomic-level stick-slip was first discovered by Mate et al., who were using AFM to measure friction between a tungsten tip and a graphite (0001) surface (Mate et al. 1987). The lateral signal exhibited stick-slip behavior with the spatial periodicity of the graphite lattice. Since then, this behavior has been observed on a wide range of materials, from soft materials like stearic acid crystals using silicon nitride tips to a diamond tip on a diamond surface. Typical atomic-level stick-slip behavior is illustrated in Fig. 1. This example shows the lateral force experienced by a silicon nitride tip sliding from left to right on the (0001) surface of muscovite mica. In the image, one sees a periodic lattice. The line trace shows that the lateral force starts from zero and builds up to a maximum value. During this phase of the measurement, the tip is sticking to the surface with no relative slip (although there may be some lateral deformation of the tip and sample). The arrow indicates the occurrence of the first slip event. The slip involves the tip moving the equivalent of one lattice spacing along the surface. The tip then sticks again until the maximum lateral force is reached once more, and the next slip occurs, and so on. The periodicity of the slip events is 0.529 nm, which is equal to the lattice constant of the mica (0001) surface. The well-defined force at which the tip slips, F_f, is called the static friction force.

The AFM signals measured correspond primarily to changes in the slope of the end of the AFM cantilever beam, which bends or twists due to forces normal (F_z) or parallel (F_x, F_y) to the surface, as illustrated in more detail in Fig. 2. Morita et al. have carried out a systematic study of atomic-level stick-slip on a range of materials, demonstrating precise determination of the slip motions that take place (Morita et al. 1996). As seen in Figs. 2 and 3 torsional or buckling rotations at the end of cantilever occur due to frictional forces acting either transverse (F_x) or parallel (F_y), respectively, to the long axis of the cantilever's projection onto the sample. The data in Fig. 3 and



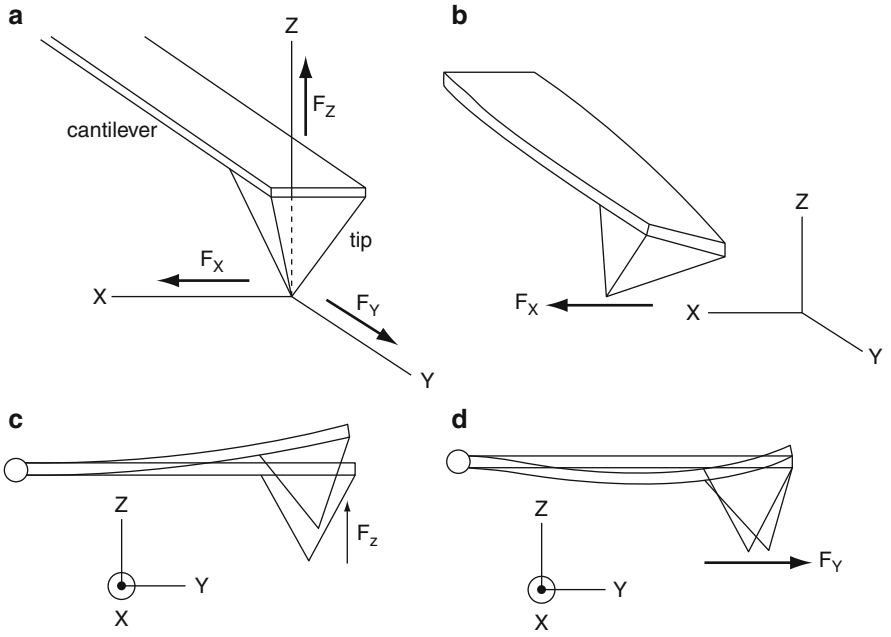
Atomic-Level Stick-Slip, Fig. 1 (**a**) Lateral force AFM image of the muscovite mica (0001) surface. Image size: (7.5 × 7.5) nm^2. The fast scan direction is from left to right. The *filled circles* represent the lattice of the mica unit, whose symmetry and periodicity (0.529 nm) coincide with the lateral forces. (**b**) Line trace of the section indicated in (**a**). The lateral force exhibits "stick–slip" behavior, where the lateral force builds up to some well-defined maximum value, and then quickly relaxes (*arrow*). During the relaxation, the tip slips by one unit cell. This behavior repeats itself with the lattice periodicity (From Szlufarska et al. 2008)

from many other experiments demonstrate that, on an ordered sample, the tip prefers to reside in positions *in registry with the sample lattice*, even though the atoms on the surface of the tip may not be ordered themselves. The importance of interfacial commensurability will be discussed further below. This periodic interaction is responsible for *all* atomic-level contrast images obtained with contact-mode AFM. One must not imagine the AFM tip smoothly tracing out atomic corrugations as with a scanning tunneling microscope (STM), but instead realize that the relative tip-sample motion is *discontinuous*.

The first few observations of this phenomenon were acquired with highly anisotropic samples, such as graphite and mica, which exhibit strong covalent bonding within each layer but weaker van der Waals or electrostatic forces between the layers. These materials cleave easily to expose their basal planes. It was suggested that the periodic forces occurred because a flake of the layered material had become attached to the tip (Morita et al. 1996). Thus, the tip and sample structures were commensurate, and a periodic interaction would be expected. However, further measurements reported stick-slip on materials that did not possess such bonding anisotropy, such as NaCl, gold, and diamond. Atomic-level stick-slip can thus occur between the sample and the tip itself.

Stick–Slip and Contact Size

Initially, several researchers misunderstood data like those in Fig. 1, thinking that true atomic resolution was achieved. One aspect that contributed to this misunderstanding is that there is no way to distinguish between the buckling and bending deformation modes of the cantilever (see Fig. 2c and 2d). As a result, atomic-level stick-slip behavior was misinterpreted as a topographic signal from the corrugation, as seen in STM images. But this was not the case. The lack of true atomic resolution in contact AFM can be understood in light of the contact mechanics. When the tip is in contact with a given sample, for typical tip radii, loads, and elastic constants, the contact is larger than a single atom. For example, a 20-nm radius silicon nitride tip exerting a 1-nN load on a mica sample produces a contact area involving nearly 15 mica unit cells, as estimated using the Hertz theory. Furthermore, Hertz theory neglects tip-sample adhesion, which, if included, makes the estimated contact area even larger and can ensure a substantial contact area even at the lowest possible applied loads. Atomistic models confirm this argument (Szlufarska et al. 2008). As a result, contact-mode AFM cannot have single-atom resolution as an STM does. In fact, Mate's original paper presented similar calculations that showed the contact area to be far greater than a single atom contact (Mate et al. 1987). This observation has several consequences; for example, the lateral resolution of features is limited by the contact area, and, as a result, point defects are not imaged. It therefore remains to be explained why, despite having a multiple atom contact and (most likely) a non-commensurate tip structure, the interaction between the tip and sample possesses the periodicity of the sample's atomic lattice.



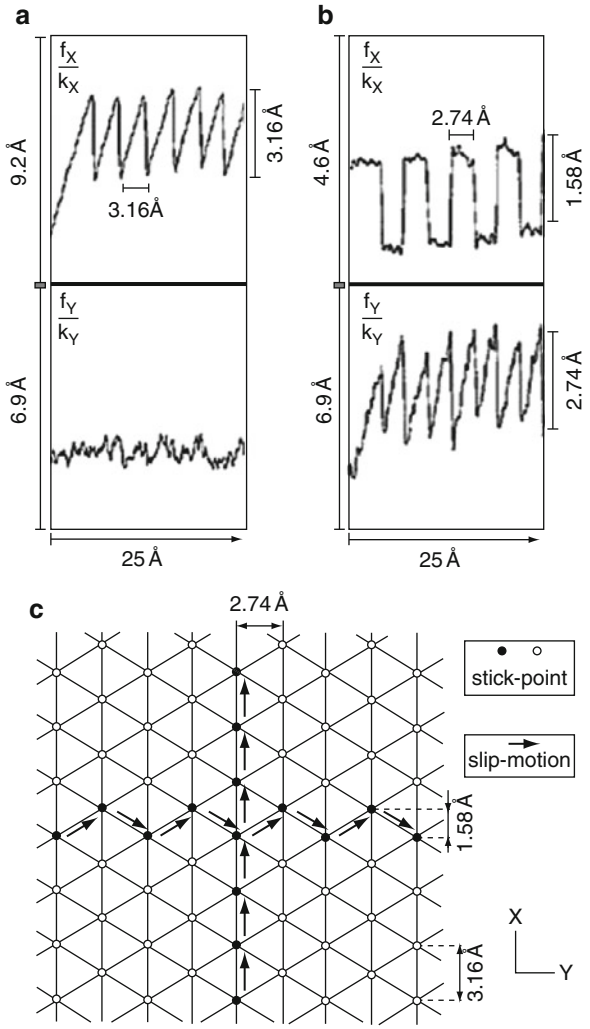
Atomic-Level Stick-Slip, Fig. 2 (**a**) Force components that act on the tip apex of an AFM cantilever. F_x, F_y, and F_z are the forces across, along, and normal to the cantilever, respectively. These forces cause torsion (**b**), bending (**c**), and buckling (**d**), respectively (From Morita et al. 1996)

Comparison with Conventional Stick–Slip Motion

The term *stick-slip* must be used carefully. Historically, *stick-slip* refers to macroscopic behavior involving multiple contacting asperities (Mate 2008). A creaking door hinge, a bowed violin string, screeching tires, and earthquakes are all examples of macroscopic stick-slip. Furthermore, stick-slip in micrometer-scale single-asperity contacts has been observed frequently using the surface forces apparatus. A rich variety of phenomena are involved in these examples, but the unifying principle is that the instability results from the dependence of friction upon the interfacial sliding speed or the static contact time in combination with the presence of some elastic compliance in the system. Specifically, if friction during sliding is lower than in the static case, if sliding friction reduces with increasing sliding speed, or if friction grows with time in static contact, then stick-slip instabilities can result (Mate 2008).

Consider a force applied through an elastic spring of a given stiffness to a slab of material in static contact with another material. The remote end of the spring is moved at a fixed pulling speed. Initially, because the surfaces are stuck together, the spring stretches and thus the lateral force the spring exerts on the slab increases. Once this force exceeds the static friction force, sliding begins. If friction is lower at higher interfacial sliding speeds, then this leads to increasingly faster relaxation of the spring force until it is no longer large enough to maintain sliding (i.e., it falls below the kinetic friction force for that relative sliding speed). The system then sticks again and the cycle repeats. This behavior is influenced by factors including the surface roughness and sliding speed-dependent effects particularly evident in viscous or viscoelastic materials. In contrast, in *atomic-level stick-slip*, the interface is atomically smooth, wear does not occur, and the contact may involve only solid, largely elastic materials, although the behavior is also seen in viscoelastic materials. No strengthening of static friction with contact time or decrease in sliding friction with sliding speed is required.

Consequently, the dependence of atomic-level stick-slip friction on pulling speed deviates from that typically seen for macroscopic stick-slip behavior. Macroscopic interfaces typically exhibit decreased static friction with increasing pulling speed as a result of the dependence of friction on interfacial contact time mentioned above.



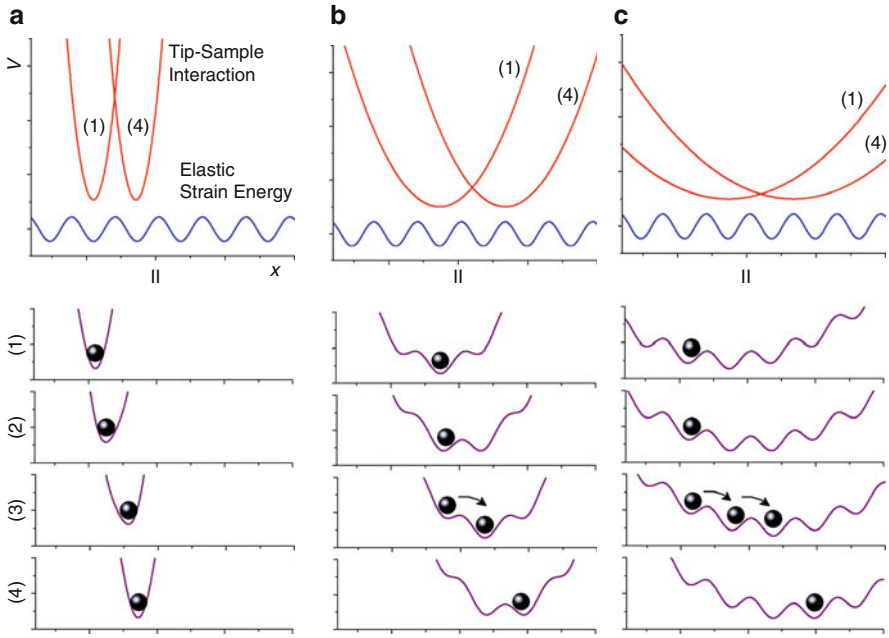
Atomic-Level Stick-Slip, Fig. 3 Friction data for lateral cantilever twisting (f_x/k_x) and buckling (f_y/k_y) due to frictional forces parallel to a MoS_2 surface acting on a Si_3N_4 tip, where f_x, f_y and k_x, k_y are the friction force along x and y and the torsional and flexural stiffness of the cantilever, respectively. In (**a**), the cantilever was scanned along the *x* direction indicated in (**c**) (i.e., perpendicular to its long axis). Stick–slip behavior resulted in periodic lateral twisting of the cantilever and no appreciable back-and-forth longitudinal buckling of the cantilever. In (**b**), the cantilever was scanned along the *y* direction indicated in (**c**) (i.e., parallel to its long axis). This time, the cantilever buckled back and forth as it stuck (*top*) and twisted back and forth (*bottom*) while scanning. Individual stick points and the path of the tip were mapped out, as indicated in (**c**), which corresponded to the MoS_2 unit cell (From Morita et al. 1996)

In contrast, a near-logarithmic *increase* of the static friction force with increasing pulling speed is attributed to reaction rate-limited slipping from one lattice site (or potential energy minimum) to the next (Gnecco et al. 2000). The atomic-scale stick-slip process can thus be viewed as a series of chemical reactions with potential energy barriers that limit motion, where the slip event is the "reaction." Tip atoms in the "stuck" position try to surmount these barriers at a natural attempt frequency. The slower the pulling speed, the more likely it is that they will be able to slip out of their local potential minima at a lower applied lateral force. If the tip is pulled at a high speed, the likelihood that these natural escape attempts will contribute to the slip process is reduced, and the force required to counteract friction will be greater than if the tip atoms were pulled along more slowly. Observations of atomic-scale static friction decreasing with pulling speed have also been reported, but this humidity-dependent behavior is observed to arise from the formation of a water meniscus between the tip and certain surfaces (Greiner et al. 2012; Riedo et al. 2002). As mentioned in the theory section below, the precise explanation for the speed and temperature dependence is a matter of ongoing debate.

Commensurability at the Interface

Interfacial commensurability, i.e., periodic coincidence of the lattices, is not a necessary condition for the occurrence of stick-slip. For example, atomic-level stick-slip has been observed with AFM tips made out of amorphous silicon nitride and oxide (Morita et al. 1996). Even if the tip atoms are ordered, they will not necessarily be in an arrangement that is commensurate with the sample's lattice. Without a commensurate interface, a sufficiently large tip would have no preferred relative positions in which to reside, and therefore smooth sliding could be expected. In light of this argument, the regular appearance of atomic-level stick-slip for a wide range of tip materials and sizes is surprising. Not only does this phenomenon occur for many tip and sample materials, but it has also been observed in humid and dry air, liquid and vacuum, and from cryogenic to elevated temperatures.

While atomic-level stick-slip is *often* observed with crystalline samples, it is not *always* observed. Conditions can vary so that a given tip can alternate between producing stick-slip motion and not. This behavior is the topic of much discussion among experimentalists, but no systematic study of the specific conditions that govern the occurrence of stick-slip friction has been carried out.



Atomic-Level Stick-Slip, Fig. 4 Energy V versus displacement x described for the 1-D Prandtl-Tomlinson model for (**a**) continuous sliding, (**b**) stick-single slip, and (**c**) stick-double slip. *Upper* plots illustrate the separate energetic contributions from the periodic tip-sample interaction potential (the lower sinusoidal curves) and the strain energy in the elastic components of the system (the upper parabolic curves). The *lower* plots show the total energy changing as the scan progresses, with a sphere indicating the location of a monatomic tip

Theoretical Approaches

Several theoretical efforts to explain and model atomic-level stick-slip behavior, specifically in the context of force microscopy, have appeared in the literature. These studies can be divided into analytical approaches and molecular dynamics (MD) simulations. The analytical approaches primarily address the *mechanics* of stick-slip behavior, i.e., a potential energy distribution (frequently referred to simply as a potential) is assumed and the resulting behavior studied. Most of the analytical approaches build on the Prandtl-Tomlinson model (Tomlinson 1929; Prandtl 1928; Popov and Gray 2012). Some of these models represent the tip as a single atom or a single entity without internal degrees of freedom, although multi-atom (tens of atoms) tips have also been considered (Gyalog and Thomas 1997; Weiss and Elmer 1996). Scanning is simulated by increasing the lateral displacement between the support and the sample. The tip initially resides in a potential minimum that is determined by the tip-sample interaction. Because finite static friction due to tip-sample interactions inhibits sliding of the tip, elastic energy is built up in the cantilever and in elastic deformations of the tip and sample themselves.

The total energy of the system is comprised of the interaction energy and the elastic energy stored in the cantilever and the deformed contact (Fig. 4). If the (lateral, or torsional) spring is compliant enough compared with the corrugation of the potential energy, a critical point is eventually reached, where the elastic strain energy becomes sufficient to move the system out of the potential minimum. As a result, slip between tip and sample takes place. In this slip stage, the cantilever and the contact quickly relax, the previously stored energy is released, and the motion is brought to a stop as the tip finds a new potential minimum, the closest one being one unit cell over. This stick-slip motion generates vibrations both in the sample and the cantilever. The phonons excited in this process carry energy away from the interaction region. Since phonon group velocities are much higher than typical AFM tip scanning speeds (even when slipping), this relaxation occurs before the next stick event. The collective results of the analytical models can be summarized as follows:

1. The stick-slip instability can be interpreted as the system (tip and sample) residing in or searching for

potential energy minima, where the energy is the sum of the tip-sample interaction and elastic energy stored in the torsion of the cantilever and the lateral deformation of the contact.
2. Sufficiently small stiffness values of the cantilever springs and the contact itself and a sufficiently strong tip-sample interaction are required to produce the stick-slip instability. If this is not the case, then the stick-slip instability can be prevented, and near-frictionless sliding can occur (McClelland and Glosli 1992). This phenomenon is discussed in more detail in the next section.
3. The energy stored and then dissipated will be distributed amongst the cantilever, the tip, and the contact, depending on their relative (lateral) stiffness values and damping constants.
4. Friction decreases with increasing temperature due to thermally activated hopping across potential barriers. Increased scanning (lateral pulling) speeds will lead to increased friction because of the reduced amount of time given to allow thermally assisted sliding to occur. However, the attempt frequency and precise temperature and speed dependencies are matters of debate.
5. The entire system, which involves the tip-sample interaction, the lateral contact stiffness, and the cantilever's torsional stiffness, is non-linear in nature. The resulting dynamics can be chaotic depending on the lateral pulling speed and the tip-sample interaction.

While these insights are clearly important, they do not provide any information on the details of vibrations (energy dissipation) in the contact zone, the physical origin of the interaction forces, or the possibility of relaxation and displacement of tip atoms in the contact. Some further insight into the origins of stick-slip behavior has been provided by molecular dynamics (MD) simulations. MD simulations have revealed that stick-slip can vary with applied load, scan speed, and scan direction with respect to crystallographic directions. Slip has been shown to occur for some systems via a dislocation mechanism, whereby tip atoms that initially reside in surface *fcc* positions relieve lateral strain by shifting to *hcp* sites (Sorensen et al. 1996). The slipped and unslipped atoms are separated by a dislocation that propagates through the contact.

Interpretations of MD data must be carried out with caution, as MD approaches suffer from several significant limitations. These are primarily the following:

1. In most cases, because of computational limits, the modeled tip is approximately ten times smaller than those used in AFM experiments.
2. Also, because of computational limits, scanning speeds are several orders of magnitude faster than what is achieved in AFM experiments. Typical MD simulation speeds are 10^{-1} m/s to 10^{2} m/s versus typical AFM experimental speeds of 10^{-7} m/s to 10^{-5} m/s.
3. Simplifying assumptions are often made regarding the interaction potentials, including, in some cases, the use of a very generic Leonard-Jones potential.

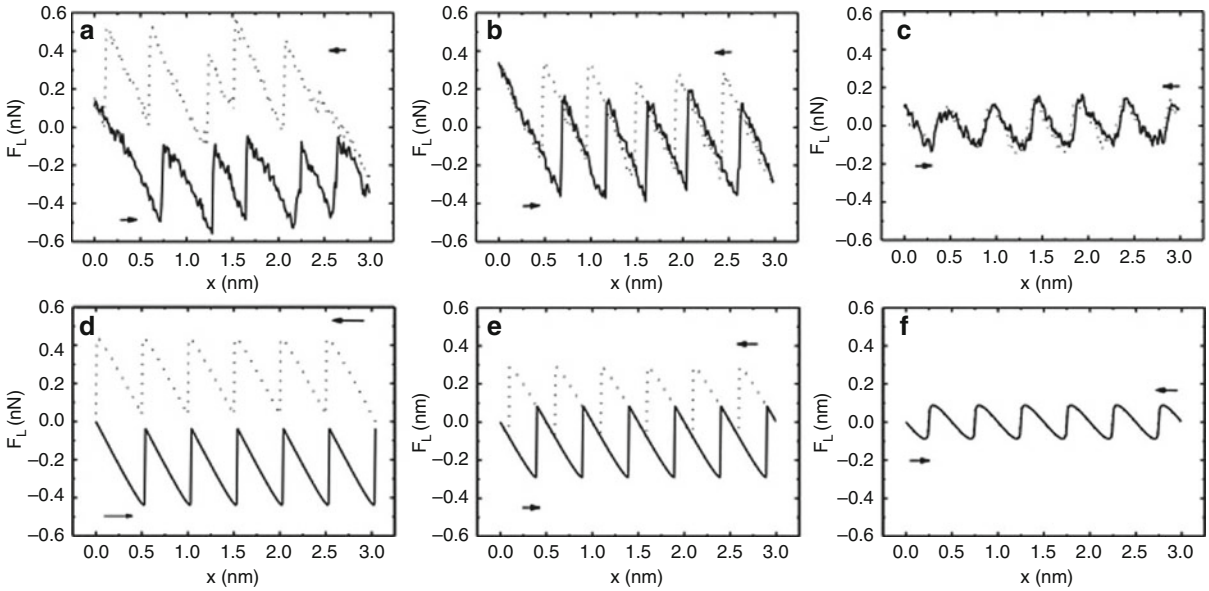
New approaches, such as multi-million atom simulations, coarse-grained approaches, or accelerated MD (Li et al. 2011), are beginning to address these issues. In addition, despite these limitations, MD provides many insights, including atomic-level descriptions of vibrational motion and energy dissipation mechanisms active during stick-slip motion, revealing that excitations are indeed highly localized in the contact zone.

Stick–Slip Transitions: Superlubricity and Multiple Slip

Recently, Socoliuc et al. used AFM to observe the existence of smooth sliding with no stick-slip when the load was sufficiently low, corresponding to extremely low energy dissipation (Socoliuc et al. 2004). As the load increased, a transition to stick-slip behavior occurred (Fig. 5). This can be understood in the context of the Prandtl-Tomlinson model. As shown in Fig. 4a, stick-slip instabilities no longer occur when the surface corrugation is sufficiently weak or the cantilever spring is sufficiently stiff. Specifically, the Tomlinson parameter $3\gamma_T = (2\pi^2 V_0)/(k_{exp}\, a_I^2)$ describes the relation between the lateral corrugation of the tip–substrate interaction V_0, the substrate lattice parameter, a_I, and the experimental lateral stiffness of the system, k_{exp}. The experimental lateral stiffness can be evaluated from the expression $\frac{1}{k_{exp}} = \frac{1}{k_{lever}} + \frac{1}{k_{tip}} + \frac{1}{k_{cont}}$, which includes the effect of the lateral stiffness of the cantilever k_{lever}, tip structure k_{tip}, and tip-sample contact k_{cont}.

Atomic stick-slip behavior is observed only if $\gamma_T > 1$, i.e., when the system is sufficiently compliant or the interfacial corrugation is sufficiently strong. When $\gamma_T < 1$, sliding occurs without stick-slip instabilities. This phenomenon has been termed "superlubricity." The term is somewhat misleading, as there can still be dynamic dissipation. So far, however, the friction force observed in these cases has been lower than the detectable limit of the AFMs used, and correspondingly the friction loops have no observable hysteresis within the experimental uncertainty.

Superlubricity has been accomplished in a variety of ways. For example, Dienwiebel et al. observed superlubricity for a graphite flake attached to the tip

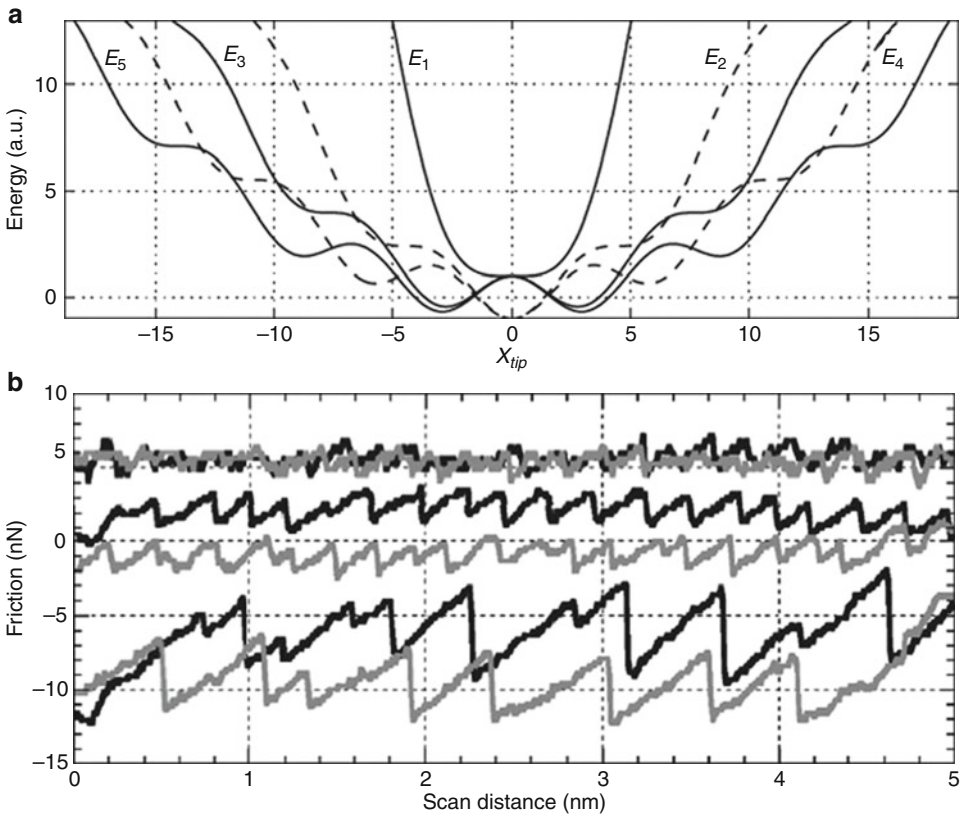


Atomic-Level Stick-Slip, Fig. 5 (**a**)–(**c**) Experimental friction loops showing the lateral force acting on the tip sliding from right to left (*dotted line*) and then left to right (*solid line*) in the (100) direction on the NaCl (001) surface in UHV. The externally applied load was (**a**) 4.7 nN, (**b**) 3.3 nN, and (**c**) 0.47 nN. (**d**)–(**f**) Corresponding numerical results from the Tomlinson model for (**d**) $\gamma_T = 5$, (**e**) $\gamma_T = 3$, and (**f**) $\gamma_T = 1$. The lateral stiffness for the calculation is chosen to be $k_{exp} = 1$ N/m and the lattice constant $a_l = 0.5$ nm. When $\gamma_T \leq 1$, smooth sliding is seen and hysteresis between the forward and the backward lateral forces disappears (i.e., there is no dissipation) (From Socoliuc et al. 2004)

sliding on a graphite surface (Dienwiebel et al. 2004). Superlubricity was only observed when the flake and the surface were oriented to be incommensurate, in agreement with the idea that incommensurability renders the corrugation of the interfacial potential sufficiently weak to allow the suppression of stick-slip instabilities. Superlubricity has also been observed for sharp AFM tips sliding over atomically flat surfaces at sufficiently low loads (Socoliuc et al. 2004; Medyanik et al. 2006). In these experiments, superlubricity was enabled by sufficiently low interfacial corrugation resulting from low applied loads (Socoliuc et al. 2004). Socoliuc et al. correlated this superlubric behavior with low values of the interfacial lateral stiffness. The demonstrated agreement with the predictions of the Tomlinson model is impressive. Recently, methods have been described that can be readily applied to practical situations. By oscillating the tip or sample at frequencies corresponding to normal resonances of the system, one can intermittently achieve low loads and thus low interfacial corrugations with very little actuation required. Such an oscillating tip is able to slide stably at the instants where the corrugation (total load) is low, resulting in a lower net friction. This same technique has been applied to macroscopic systems.

Another type of transition was predicted by Johnson and Woodhouse, who showed that under certain conditions slip may occur over an integer number of lattice spacings (Johnson and Woodhouse 1998). This phenomenon is called *multiple slip*. In fact, multiple slip was observed in the original letter reporting atomic-level stick-slip by Mate et al. (1987) but since then it has been rarely discussed until recently (Medyanik et al. 2006; Roth et al. 2010). Johnson and Woodhouse identified the relationships between the lateral (i.e., torsional) cantilever stiffness, the lateral stiffness of the elastically deformed contact itself, and the corrugation of the lateral force interaction as key parameters controlling the transition to multiple slips (Johnson and Woodhouse 1998). An adjustable damping factor was introduced that represents the dynamic energy dissipation in the tip or sample materials or in the cantilever itself. The transition from single to double slips occurs when high-frequency fluctuations in the lateral force, triggered by the slip instability, overshoot the corrugated lateral tip-sample interaction force. The possibility of an overshoot reduces with increased damping.

Analytical approaches have been employed to describe transitions from single to multiple slip in atomic-scale



Atomic-Level Stick-Slip, Fig. 6 (**a**) Calculation of the energy landscapes (energy vs. position of the tip) corresponding to five different values of lateral contact stiffness. For each energy curve E_i (except $i = 1$) there are i local minima that correspond to stable equilibrium states of the system. Therefore, as the system becomes unstable due to the motion of the cantilever, there are i possible destinations in which to slip. (**b**) Experimental friction behavior on the (0001) surface of HOPG exhibits transition to multiple slips: Smooth sliding is observed at the lowest load (*top*), single slip is observed at intermediate loads (*middle*), and mostly double slips occur at the largest load (*bottom*) (From Medyanik et al. 2006)

friction (Fig. 4). For example, Conley et al. used a combination of numerical methods to deal with issues of the complex dynamics in atomic-scale friction (Conley et al. 2005). The authors considered a quasistatic limit and transitions between multiple slip modes by solving the equation of motion numerically. An analytical solution of the transition between different slip regimes for the simple case of the one-dimensional Tomlinson model in the quasistatic limit was recently reported by Medyanik et al. (2006). The authors analyzed the energy landscape and showed how the number of local energy minima increases with increasing interfacial corrugation (Fig. 6a). Single and multiple slips correspond to sliding of the tip to the nearest, the next nearest, or the next-next nearest, etc., local minimum. Slipping to further minima can occur only with sufficiently low energy dissipation during slip.

Transition from single to multiple slip occurs with increasing load, which indicates that corrugation increases with load. Specifically, the existence of multiple slip regimes is governed by the Tomlinson parameter, γ_T, reaching characteristic values. In other words, $\gamma_T = 1$ represents the transition from smooth sliding to slipping by one lattice site. The possibility of slips of higher multiplicity occurs for larger critical values of γ_T. In the same paper (Medyanik et al. 2006), the authors reported experimental observation of the dependence of stick-slip behavior on load. The experiments were performed on a highly oriented pyrolitic graphite (HOPG) sample, and, as shown in Fig. 6b, the system exhibited superlubricity at the lowest applied load. At higher loads, stick-slip instabilities occurred with the periodicity of the HOPG lattice, while increasing the load even further leads to slips over integer multiples of the lattice spacing, as predicted by the model.

Remaining Questions

There has not yet been any clear conclusion indicating under exactly what conditions stick-slip behavior occurs. Often, images like the one shown in Fig. 1 are *not* obtained. It is possible that under the same loads, with the same sample and with the same cantilever, some unknown change in the tip occurs and stick-slip is suddenly observed. The reasons for this are not established. Furthermore, no one has studied whether friction varies with load in the same manner in the presence and absence of stick-slip.

Another unresolved question pertains to stick-slip periodicity. Most accounts so far report *one* stick-slip event per surface unit cell, even when the unit cell contains more than one atomic species, such as the surface of an alkali halide crystal. One exception is the large unit cell of Si(111)-77, as measured in ultrahigh vacuum (UHV) with tips coated with polytetrafluoroethylene, where multiple stick-slip events per unit cell were resolved (Howald et al. 1995).

As discussed above, energy released during sliding is carried away by phonons excited in the sliding process. Phonon frequencies are eleven orders of magnitude higher than AFM scanning frequencies, and the relevant dissipation processes occur quickly. It has been found recently in experiments involving Si tips on KBr samples that slip times in atomic stick-slip can be as long as 10 ms. Such time scales are currently inaccessible to conventional atomistic simulations. However, new accelerated simulation techniques, or the use of more powerful computers and efficient algorithms, may provide routes to addressing this challenge.

Key Applications

While macroscale stick-slip processes are capable of producing tangible end products, from music to mountain ranges, atomic-scale stick-slip has yet to be incorporated in any direct application. Stick–slip is used by microelectromechanical systems for precision applications, demonstrating positioning down to 10 nm (de Boer et al. 2004). It is conceivable that stick-slip methods could be extended to atomic-scale positioning by integrating nanoscale contacts into a device, such that atomic-level stick-slip at the contact point can be used as a precise and accurate indicator of position. Nonetheless, such applications have not yet been realized in a device. At the moment, the most commonly used application for atomic-level stick-slip is the nanoscale lateral spatial calibration of AFMs. This procedure is one of a few and perhaps the fastest of the ways to calibrate an AFM at this scale.

References

W.G. Conley, A. Raman, C.M. Krousgrill, Nonlinear dynamics in Tomlinson's model for atomic-scale friction and friction force microscopy. J. Appl. Phys. **98**, 053519 (2005)

M.P. de Boer, D.L. Luck, W.R. Ashurst, R. Maboudian, High-performance surface micro-machined inchworm actuator. J. Microelectromech. Syst. **13**, 63 (2004)

M. Dienwiebel, G.S. Verhoeven, N. Pradeep, J.W.M. Frenken, J. A. Heimberg, H.W. Zandbergen, Superlubricity of graphite. Phys. Rev. Lett. **92**, 126101 (2004)

C. Greiner, J.R. Felts, Z. Dai, W.P. King, R.W. Carpick, Controlling nanoscale friction through the competition between capillary adsorption and thermally-activated sliding. *ACS Nano*, **6**, 4305 (2012)

E. Gnecco, R. Bennewitz, T. Gyalog, C. Loppacher, M. Bammerlin, E. Meyer, H.-J. Guntherodt, Velocity dependence of atomic friction. Phys. Rev. Lett. **84**, 1172 (2000)

T. Gyalog, H. Thomas, Friction between atomically flat surfaces. Europhys. Lett. **37**, 195 (1997)

L. Howald, R. Lüthi, E. Meyer, H.-J. Güntherodt, Atomic-force microscopy on the Si(111)77 surface. Phys. Rev. B **51**, 5484 (1995)

K.L. Johnson, J. Woodhouse, Stick–slip motion in the atomic force microscope. Tribol. Lett. **5**, 155 (1998)

Q. Li, Y. Dong, D. Perez, A. Martini, R.W. Carpick, Speed dependence of atomic stick-slip friction in optimally matched experiments and molecular dynamics simulations. Phys. Rev. Lett. **106**, 126101 (2011)

C.M. Mate, *Tribology on the Small Scale: A Bottom up Approach to Friction, Lubrication, and Wear* (Oxford University Press, Oxford/New York, 2008)

C.M. Mate, G.M. McClelland, R. Erlandsson, S. Chiang, Atomic-scale friction of a tungsten tip on a graphite surface. Phys. Rev. Lett. **59**, 1942 (1987)

G.M. McClelland, J.N. Glosli, in *Fundamentals of Friction*, ed. by I.L. Singer, H.M. Pollock. (Kluwer, Dordrecht; 1992), p. 405

S. Medyanik, W. K. Liu, I.-H. Sung, R.W. Carpick, Predictions and observations of multiple slip modes in atomic-scale friction. Phys. Rev. Lett. **97**, 136106 (2006)

S. Morita, S. Fujisawa, Y. Sugawara, Spatially quantized friction with a lattice periodicity. Surf. Sci. Rep. **23**, 3 (1996)

V. L. Popov, J.A.T. Gray, Prandtl-Tomlinson model: history and applications in friction, plasticity, and nanotechnologies. Z. Angew. Math. Mech. **92**(9), 683–708 (2012)

L. Prandtl, A conceptual model to the kinetic theory of solid bodies. Z. Angew. Math. Mech. **8**, 85 (1928)

E. Riedo, F. Levy, H. Brune, Kinetics of capillary condensation in nanoscopic sliding friction. Phys. Rev. Lett. **88**, 185505 (2002)

R. Roth, T. Glatzel, P. Steiner, E. Gnecco, A. Baratoff, E. Meyer, Multiple slips in atomic-scale friction: an indicator for the lateral contact damping. Tribol. Lett. **39**, 63 (2010)

A. Socoliuc, R. Bennewitz, E. Gnecco, E. Meyer, Transition from stick-slip to continuous sliding in atomic friction: entering a New regime of ultralow friction. Phys. Rev. Lett. **92**, 134301 (2004)

M.R. Sørensen, K.W. Jacobsen, P. Stoltze, Simulations of atomic-scale sliding friction. Phys. Rev. B **53**, 2101 (1996)

I. Szlufarska, M. Chandross, R.W. Carpick, Recent advances in single-asperity nanotribology. J. Phys. D: Appl. Phys. **41**, 123001 (2008)

G.A. Tomlinson, A molecular theory of friction. Philos. Mag. Ser. **7**, 905 (1929)

M. Weiss, F.-J. Elmer, Dry friction in the Frenkel-Kontorova-Tomlinson model: static properties. Phys. Rev. B. **53**, 7539 (1996)

Atomic-Scale Stick-Slip

► Atomic-Level Stick-Slip

Attractive Interaction Force

► Adhesion in the Animal World

Attrition

► The Tribology of Dental Materials

Auger Electron Spectroscopy (AES)

Thomas Schuelke
Fraunhofer Center for Coatings and Laser Applications, Fraunhofer USA, Inc., Michigan State University, East Lansing, MI, USA

Definition

AES is an analytical technique that involves probing a sample with an energetic particle beam to generate Auger electrons, which are emitted from the sample atoms with element-specific kinetic energies. In AES the kinetic energy distribution of these emitted electrons is analyzed to identify chemical elements, their composition, and distribution in the sample and to study the chemical environment of the emitting atoms. The typical application of AES involves a probing beam of electrons with primary kinetic energies of several keV. This electron beam is focused typically onto a solid-state surface, exciting its immediate surrounding region with a spatial resolution of smaller than 100 nm. The technique is very surface sensitive with an information depth of less than 5 nm, or a few atomic layers. Surface adsorbate quantities of less than 1% of an atomic monolayer can be detected. Qualitative, and to a limited degree quantitative, information can be derived with respect to the chemical bonding state of surface atoms. AES requires ultra-high vacuum equipment operating at a base pressure of lower than 10^{-7} Pa. To obtain complementary information the method can be beneficially combined with other surface characterization techniques such as photoelectron spectroscopy (XPS), secondary ion mass spectroscopy (SIMS), or low energy electron defraction (LEED). In combination with an ion beam source, AES can provide highly resolved three-dimensional chemical composition information through concentration depth profiling.

Scientific Fundamentals

Mechanism

Auger Effect

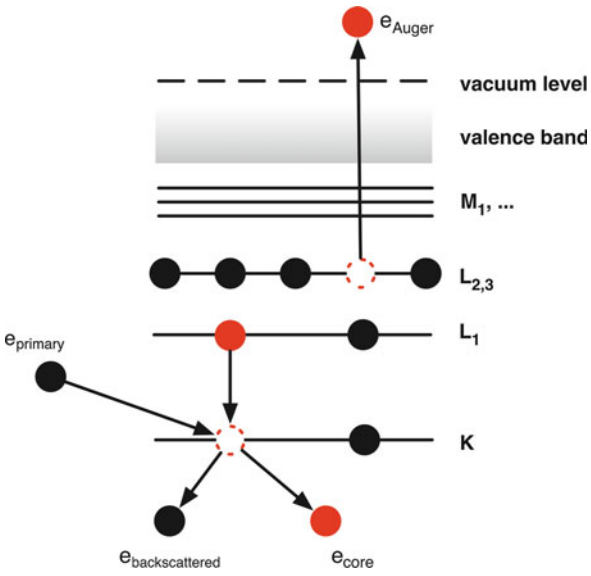
The Auger effect (named after Pierre Auger, 1899–1993) is a radiationless atomic de-excitation mechanism involving three steps. Figure 1 schematically shows an example of a specific $KL_1L_{2,3}$ Auger transition involving three atom shell energy levels (K, L_1, and $L_{2,3}$) and their shell electrons (here marked in red). First a primary probing beam electron $e_{primary}$(i.e., 3–30 keV) removes a core-level shell electron e_{core} from a specimen atom by impact ionization and leaves the shell region as an inelastically backscattered electron $e_{backscattered}$. An outer shell electron of the same atom then fills the generated core level hole (i.e., transitions from the L_1 to the K shell in this particular example). This internal transition frees energy that couples to a second electron of a third shell (here $L_{2,3}$) in the same atom, the *Auger electron*, which ejects from the atom if the transferred energy exceeds its binding energy. The overall process leaves a two-hole final state, which in the discussed example has ionized L_1 and $L_{2,3}$ shells.

As described in this example, the Auger effect involves three shell electrons of the same atom. Consequently, the lightest detectable chemical element is lithium. Hydrogen and helium do not have Auger transitions due to an insufficient number of shell electrons. All other elements have at least one Auger transition yielding Auger electrons with a kinetic energy below 2 keV.

AES Energy Level Notations

The energy levels in Auger electron spectroscopy are labeled in X-ray notation (Briggs and Grant 2003). The principal quantum numbers (e.g., 1, 2, 3, etc.) are referred to as K, L, M, etc. Subscripted numbers (e.g., $L_1 \ldots L_3$, $M_1 \ldots M_5$, etc.) denote the combinations of orbital angular momentum and electron spins. If the energy levels are too close to resolve, they are designated with a comma between subscripts (e.g., $L_{2,3}$).

In this nomenclature, the Auger process shown in Fig. 1 represents a $KL_1L_{2,3}$ transition. When applying AES to analyze solid-state samples, the outer shell-level electrons contribute to chemical bonding and their orbitals overlap forming the valence band. The letter V is



Auger Electron Spectroscopy (AES), Fig. 1 Schematic illustration of a typical Auger process for the case of a $KL_1L_{2,3}$ transition with electron beam probing

often used to identify these Auger transitions (e.g., LVV). Sometimes the letter C is used as a generic reference to non-bonding core levels (e.g., CVV). If the Auger relaxation transition involves an electron from the same principal shell that was originally ionized (e.g., L_1L_2M) it is often described as a Coster-Kronig transition.

The Kinetic Energy and Intensities of Auger Electrons

The kinetic energy of the Auger electron is independent of the mechanism or particle type that generates the initial core-level hole. Auger electrons can be generated from solid, liquid, and gaseous samples. The further discussion assumes the typical application of using an electron beam in an ultra-high vacuum system to probe surfaces of solid samples.

Due to the specific nature of atom shell energy levels, the kinetic energy of the Auger electron is characteristic to the chemical element from which it originates. For example, for a $KL_1L_{2,3}$ transition the kinetic energy of the Auger electron is approximately:

$$E_{kinetic}^{Auger} \approx E_{binding}^{K} - E_{binding}^{L_1} - E_{binding}^{L_{2,3}}$$

$E_{kinetic}^{Auger}$ ... kineti energy of the emitted Auger electron

$E_{binding}^{K}$ ... K shell electron binding energy in neutral atom

$E_{binding}^{L_1}$ ... L_1 shell electron binding energy in neutral atom

$E_{binding}^{L_{2,3}}$ ... $L_{2,3}$ shell electron binding energy in neutral atom

(1)

Equation 1 is an approximation and the right side requires a correction term, which accounts for changed atom shell energy levels as a consequence of the ionization state of that atom (one hole and two hole binding states). The binding energy of a given shell in a single ionized ion is slightly larger than that of its neutral state. The correction term energy is on the order of 10 eV. Auger electrons are typically detected in the energy range from 50 to 3,000 eV. Binding energies and relaxation mechanisms also depend on the chemical environment. Relative shifts of the kinetic energy of Auger electrons on the order of a few eV provide information about changes in the chemical environment, as they would occur during surface chemical reactions.

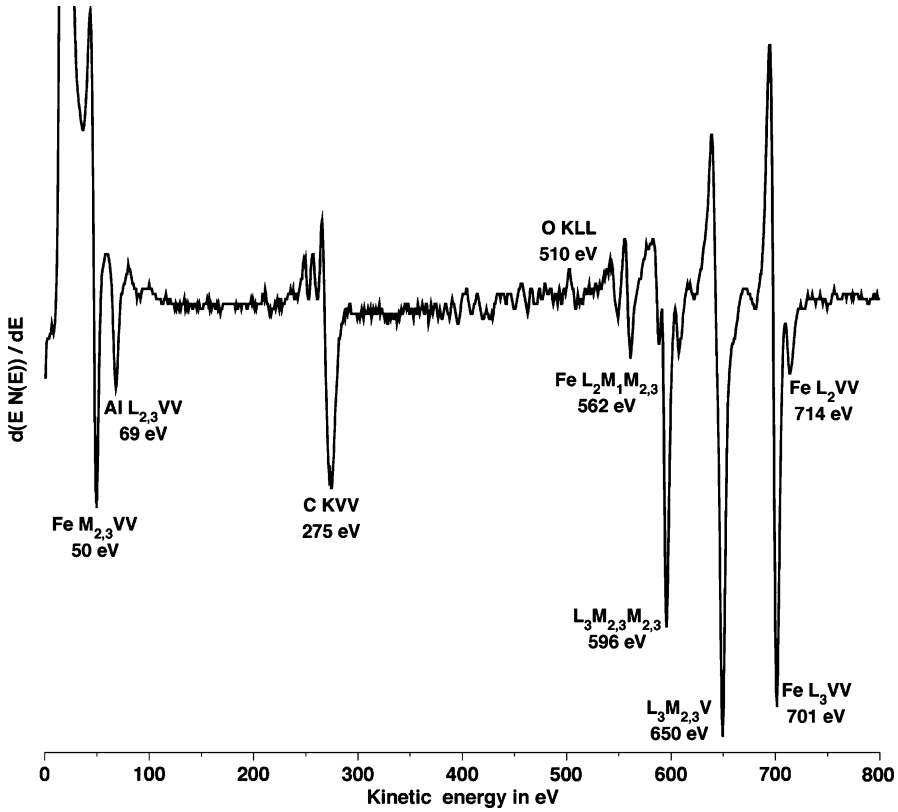
Characteristic changes of line shapes in the Auger spectrum can be observed if valence electrons are involved in Auger transitions. These line shapes are sometimes referred to as fingerprints, which in many cases allow the quick identification of the bonding state of certain surface atoms. The semi-quantitative interpretation of these line shapes is well developed (Ramaker 1991). For a CVV transition involving two valence states, the observed line shape reflects the self-convolution of the density of states in the valence band. Although experimental and theoretical agreement is seldom perfect, the observed trends reproducibly reflect the physical and chemical conditions, making the study of line shapes based on fingerprinting a very practical method.

The intensity of the Auger electron current forms the basis for quantitative analysis, and peak heights or peak-to-peak heights in derivative spectra are interpreted for this purpose. Since the probing beam excites sample atoms to a depth of about 1 μm, there are also inelastically scattered Auger electrons contributing to the background signal. Therefore, the Auger electron current from a sample cannot be directly measured due to the simultaneous presence of a background signal.

Relating Auger electron intensities to the quantity of chemical elements is not straightforward because yields of several possible Auger transitions depend on the transition type and the atomic number Z of the chemical element. For example, the Auger de-excitation mechanism competes with X-ray fluorescence relaxation. K-level Auger transitions are more likely to occur for lighter elements than photon emission. Likewise, Coster-Kronig transitions only occur for certain ranges of Z numbers (Thompson et al. 1985).

Auger Spectrum

During the detection process the Auger electron signals appear as peaks at a characteristic energy in the $N(E)$



Auger Electron Spectroscopy (AES), Fig. 2 Derivative AES spectrum of a single crystalline iron sample alloyed with aluminum and carbon

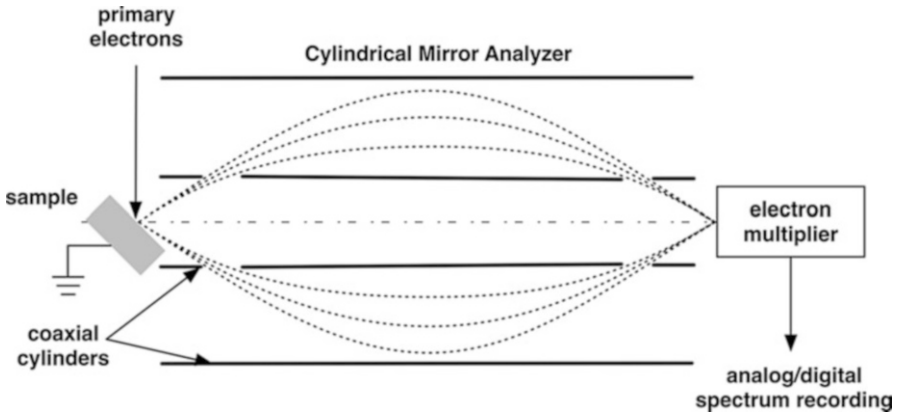
spectrum on a continuous background of secondary and backscattered electrons below the energy of the elastically scattered electrons from the incidence beam. To improve the visibility of Auger electron peaks, the spectra are plotted in a derivative form such as $dN(E)/dE$. The spectrum is frequently delivered in the form of $d(E \cdot N(E))/dE$, when the analyzer's energy resolution is energy dependent. This is, for example, the case with commonly used cylindrical mirror analyzers.

Figure 2 shows a derivative Auger spectrum from a (111) surface of a single crystalline iron sample. Prior to the measurement, the sample was annealed at 365°C for 30 min. The prominent features in this spectrum are the CVV transitions of aluminum, carbon, and iron. The primary electron beam energy is 3 keV. The various characteristic iron Auger transitions are marked in the spectrum and identified using tabulated data from the literature (Coad and Rivière 1971). The spectrum also shows LVV aluminum, CVV carbon, and KLL oxygen peaks. The shape of the carbon peak is a typical fingerprint of metal-carbon bonding (carbide).

Experiment Configuration and Instrumentation

The escape depth of the Auger electrons is related to the material and energy-dependent inelastic mean free path, which is typically less than 5 nm for the kinetic energy range of Auger electrons. In fact, for many elements the escape depth of Auger electrons with kinetic energies of up to 500 eV is less than 1 nm. Therefore, Auger electrons provide chemical information originating from a few atomic surface layers only, making the technique very surface sensitive. Generally, partial surface coverage of as little as 1% of an atomic monolayer already affects an Auger spectrum and thus can be detected.

This sensitivity has implications on sample preparation, sample handling and the required vacuum conditions. The recording of an overview (wide energy range) Auger spectrum typically takes about 100 s. The monolayer formation time of air at ambient temperature is roughly estimated to be $\tau \approx 2.5 \times 10^{-4}/p$ (unity sticking coefficient, pressure p in Pa, time τ in s) or 2.5 s at 10^{-4} Pa. Thus, ultra-high vacuum (UHV) conditions at pressures



Auger Electron Spectroscopy (AES), Fig. 3 Schematics of a cylindrical mirror energy analyzer in an AES experiment setup

of lower than 10^{-7} Pa are required to sufficiently reduce monolayer adsorption from residual gas molecules during the recording process of the Auger spectrum.

The instrumentation directly used to generate and analyze Auger electrons includes (a) an electron optical column to generate, focus, and scan the primary electron beam, (b) an energy discriminating analyzer to filter low-energy electrons as a function of their kinetic energy, and (c) an electron-multiplying signal detection unit to count, convert, and record the filtered electrons.

Modern field emission electron beam guns achieve a spatial resolution of 10 nm, which are only slightly depending on the probe current in the typical range from 10^{-11} to 10^{-8} A. However, the collection of Auger electrons requires a large working distance from the sample leading to about twice the probe size compared with high-performance scanning electron microscopes (SEM).

One of the most common energy filters is the so-called *cylindrical mirror analyzer* (CMA, see Fig. 3). The CMA consists of two concentric cylinders with an adjustable electrostatic potential between them that is varied to scan through the energy range. The energy resolution $\Delta E/E$ of the CMA is a constant defined by the geometrical parameters including the radius and the opening width for electron passage of the inner cylinder. Therefore, the number of electrons passing the CMA within a given energy interval ΔE is proportional to the energy E, which is represented in the recorded spectrum as $E \cdot N(E)$.

After energy filtering, the Auger electrons are counted by electron multipliers such as microchannel plates. Accelerated in an electric field, each electron generates via multiple collisions an avalanche on the order of 10^6 electrons, which are detected and counted as a pulse.

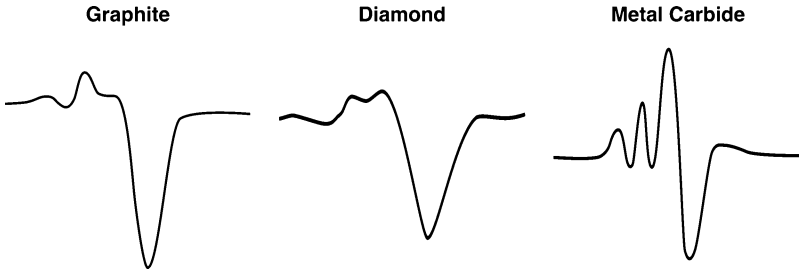
A typical Auger-equipped UHV system includes sample surface preparation instrumentation such as controlled sample heaters and the capability of ion beam treatment for the removal of surface contaminations, depth profiling, and surface charge compensation to improve the analysis of dielectric materials. For cost effectiveness and experimental efficiency, it is also common to combine multiple surface-sensitive analysis and imaging techniques in the same UHV system.

Key Applications

AES' Strengths and Shortcomings as a Surface Characterization Technique

AES is a powerful surface analysis technique with many advantageous features. The technique provides chemical information about the top-most atomic surface layers. Auger electron energies are characteristic to chemical elements. Peak shifts and shapes contain information about the chemical environment of these elements. The high-intensity electron beam excitation yields a high sensitivity, resulting in short analysis times on the order of minutes. AES has an excellent spatial resolution since the primary electron beam spot is focused to less than 100 nm, which is the biggest advantage in direct comparison to photoelectron spectroscopy (XPS). Surface mapping by beam scanning is used to generate AES images of surfaces. Three-dimensional AES mapping is possible in combination with depth profiling. Modern AES instrumentation is frequently combined with other techniques, providing complementary information.

AES also has some shortcomings. The effect of the primary electron beam needs to be considered. The high-intensity beam focused onto the surface can induce surface damage, heating, and chemical modifications such as desorption and decomposition. When studying dielectric



Auger Electron Spectroscopy (AES), Fig. 4 Carbon CVV $dN(E)/dE$ Auger fingerprints (Adopted from literature Craig et al. (1983))

materials, the electron beam exposure causes surface charging, which shifts the Auger electron energy and potentially prohibits the possibility to measure a spectrum. The quantification of AES peak heights, shapes, and shifts is theoretically difficult because of the complexity of the Auger electron excitation process. For example, it proves difficult to calculate exact Auger electron currents because they are not exclusively related to ionization cross sections. However, these shortcomings do not significantly inhibit the practical advantages and uses of AES for chemical element identification, concentration quantification based on experimental references, and chemical fingerprinting.

Chemical Element and Composition Identification

Due to its surface sensitivity, elemental specificity, and focused electron beam excitation, AES is predominantly used to provide chemical element identification and quantification with spatial resolution. The peak positions in an AES spectrum (see Fig. 2) with respect to the kinetic energy and their shapes are fingerprints of chemical elements.

In practice, there is often no need to quantify the actual current of particular Auger electrons. However, it is of great practical value to quantify the chemical composition of a particular surface in terms of percentages, which is principally related to the peak intensities usually measured as peak-to-peak heights in the derivative spectra. In addition to comparing measured spectra to reference spectra it is also common to work with so-called sensitivity factors, which are Auger peak intensities from pure reference elements ideally measured under the same conditions as the sample in question.

For example, P_x^{ref} shall denote the sensitivity factor or peak-to-peak height measured from a pure reference sample made from chemical element x and P_x shall be the peak-to-peak height of chemical element x in the test specimen. The concentration C_x of chemical element x in the test specimen is then defined by the ratio of $\frac{P_x}{P_x^{ref}}$. To avoid the need to reproduce absolute peak heights over longer periods, it is common to normalize the total composition to unity. The concentration C_x of chemical element x in the test specimen is then calculated as

$$C_x = \frac{\frac{P_x}{P_x^{ref}}}{\sum \frac{P_n}{P_n^{ref}}}, \qquad (2)$$

Where $\sum \frac{P_n}{P_n^{ref}}$ is the sum of actual to reference peak-to-peak height ratios of all detected elements n in the test specimen.

Equation 2 provides an easily applicable method to estimate quantitative element compositions. It does not take into account matrix effects (e.g., the change of Auger yields based on backscattered electrons) and may result in seriously incorrect estimates. However, within the framework of a particular study, it is a very useful approach to track, for example, the surface composition of a sample as a function of temperature (e.g., to study segregation of bulk atoms to the surface) or oxygen exposure (e.g., to study surface oxidation).

Scanning the electron beam across the specimen surface and recording Auger spectra as a function of the excitation location is used to explore the spatial distribution of chemical composition. This technique is also referred to as AES imaging. While this method appears to be straightforward, several factors have to be considered. For example, careful attention has to be paid to the data acquisition time with respect to monolayer formation times.

Depth profiling of element concentrations is possible by performing Auger scans of specific elements in combination with ion beam removal of surface layers (sputter depth profiling). Ion beam energy and impact angle determine not only the removal rate for a given substrate material but also the ion beam impact on the surface in

terms of atomic mixing. The preferred method is to use low ion energies (<1 keV) and high incidence angles (<60°) to avoid modification of the studied surface. Note that electron spectroscopy (e.g., AES, XPS) analyzes atoms that remain on the surface, whereas secondary ion mass spectroscopy (SIMS) analyzes atoms and clusters that are removed from the surface by ion beam impact.

Studying the Surface Chemical Environment

If Auger electrons originate from the valence band, the electrons are involved in chemical bonding to neighboring atoms. Additional information can be extracted based on energy shifts and modified peak shapes of CVV transitions. The Auger line shape is often used as a fingerprint of the bonding state of a chemical element.

An example is shown in Fig. 4 for the carbon CVV (KLL) transition, which significantly changes depending on the carbon bonding. The L states in this transition are the 2s and 2p valence states. Each type of carbon bond has a distinctive Auger line shape. Typical for metal-carbon bonds are two well-defined additional small peaks in the derivative spectrum. The shape is significantly altered in the case of carbon-carbon bonds. Here it is possible to distinguish between sp^2 (graphite), sp^3 (diamond), and an sp^2/sp^3 mixture (diamond-like carbon) type bonding. Peak heights might be used to calculate metal-carbon and carbon-carbon bond fractions when analyzing the peak shapes (Craig et al. 1983).

Cross-References

► Scanning Electron Microscopy (SEM)

References

D. Briggs, J.T. Grant, (Eds.). *Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*. Chichester: IM Publications and Surface Spectra Limited. ISBN 1-901019-04-7 (2003)

J.P. Coad, J.C. Rivière, The LMM auger spectra of some transition metals of the first series. Z. Phys. A Hadron. Nucl., **244**(1), 19–30 (1971)

S. Craig, G.L. Harding, R. Payling, Auger lineshape analysis of carbon bonding in sputtered metal-carbon thin films. Surf. Sci., **124**(2–3), 591–601 (1983)

D.E. Ramaker, The past, present, and future of auger line shape analysis. Crit. Rev. Solid State Mater. Sci., **17**(3), 211 (1991)

M. Thompson, M.D. Baker, A. Christie, J.G. Tyson, *Auger electron spectroscopy* (Chemical analysis series, Vol. 74). New York: Wiley. ISBN ISBN 0-471-04377-X (1985)

Austenitic Nitrocarburizing

► Nitriding and Nitrocarburizing

Automotive Lubricants

► Nanoparticles in Automotive Applications

Automotive Transmission

► Wet Clutch Friction Material: The Surfaced Groove Effect

Average Flow Model

► Average Reynolds Equation

Average Reynolds Equation

Q. Jane Wang[1], H. S. Cheng[2]
[1]Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA
[2]Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA

Synonyms

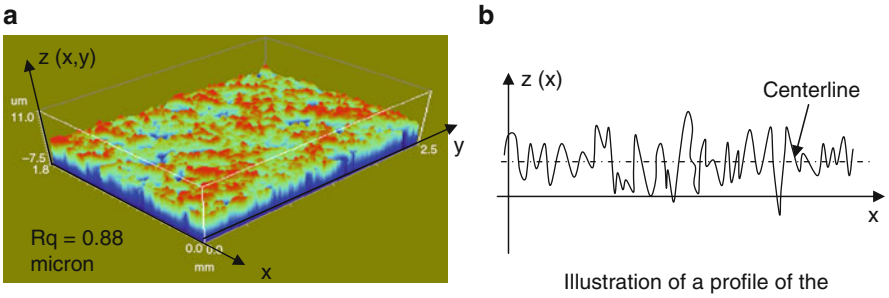
Average flow model; Effect of roughness on lubrication; Stochastic modeling of flows in lubrication

Definition

The Reynolds equation is the basic fluid dynamics relationship describing the dependence of lubricant pressure on flow film thickness. The classic Reynolds equation was derived with an assumption that the surfaces bounding the lubricating fluid are ideally smooth. However, engineering surfaces are rough and may have different scales of irregularities. When two interacting surfaces form a lubricated pair, the lubricant flow boundaries are rough surfaces. The real surfaces, not the idealized smooth surfaces, should be considered in lubrication problems. Average Reynolds equations refer to a class of modified Reynolds equations that consider the statistical nature of surface roughness in lubricant flow.

Scientific Fundamentals

Background. Engineering surfaces are complex in both material chemistry and topography. They are rough in



Average Reynolds Equation, Fig. 1 A three-dimensional rough surface and schematic of a two-dimensional roughness profile

nature, no matter how they are finished and how smooth they are polished. Surface irregularities representing the deviation of a real surface from an ideally smooth one are composed of a micro- or nanoscopic mountainous peaks and valleys. The peaks, or summits, are usually named as *asperities*; the overall surface appearance is referred to as *topography*, and the structure of surface geometry as *texture*. When two such surfaces form a tribological interface, a lubricating fluid flows across surface irregularities. If the lubricant flow film is thick, the main stream of the fluid does not "feel" surface irregularities. However, when the film thickness falls into the same order of magnitude as that of the surface roughness, notable disturbance to the flow is expected.

Mathematical descriptions of the stochastic effect of roughness on lubricant flow result in stochastic flow models, which lead to average Reynolds equations. The pioneering work by Christensen (1970) for one-dimensional problems and by Patir and Cheng (1978) for two-dimensional problems established the basic concept of the average Reynolds equation and average flow analyses. The model by Harp and Salant (2001) further takes into account the cavitation effect, and the research by Letalleur et al. (2002) and Sahlin (2008) extended the average flow concept to more general flow terms. Many have contributed in the research of the theories of rough surface lubrication, and the resultant average Reynolds equations largely fall into two types based on the characteristics of the flow factors used for Reynolds equation modification, namely the average Reynolds equations with scalar flow factors for two orthogonal flows (Patir and Cheng 1978, Hap and Salant 2001) and those with flow factor tensors including mixed flow terms as well (Letalleur et al. 2002). However, the main focus here is on Patir and Cheng's average flow model, aiming at a basic explanation of the concept.

Surface Roughness

Figure 1 (a) shows a digitized rough surface of a machine element, by means of the white-light interferometry technology, with three-dimensional topography, whose height can be expressed as $z\,(x,y)$. However, the mathematical description of a rough surface, surface statistic concept and parameters, may be more directly illustrated by a two-dimensional profile, Fig. 1 (b), where the surface height, $z\,(x)$, is taken along a traverse distance, x.

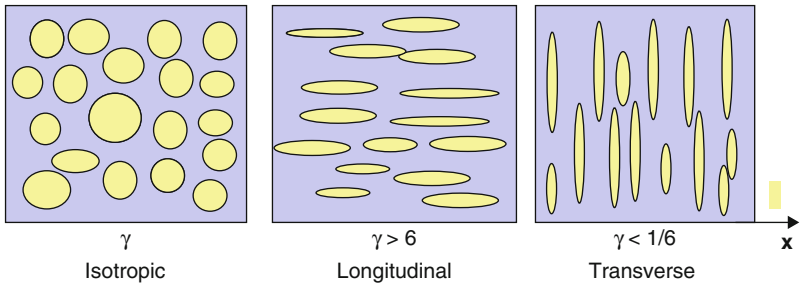
Several statistical parameters may be used in the derivation of the average Reynolds equation. The probability density function of surface height distribution is defined as height frequency count divided by the height data interval. For a profile of N surface height data, the mean, μ, or mathematical expectation, determines the centerline (Fig. 1b) of surface irregularities, or where the "ideally smooth" surface should be mathematically defined,

$$\mu = average(z) = \frac{1}{N}\sum_{1}^{N} z_i \tag{1}$$

Note that the mean is different from the centerline average, R_a, defined as the absolute deviation from the centerline, μ. The standard deviation, R_q, or the root mean square (RMS) roughness, is a commonly used parameter to characterize surface roughness. It is defined as

$$R_q = \sqrt{E(z-\mu)^2} = \left(\frac{1}{N}\sum_{1}^{N}(z_i-\mu)^2\right)^{1/2} \tag{2}$$

where E means mathematical expectation. The autocorrelation function is a spatial structural function commonly used to reveal the structural information of a surface. Mathematically, the autocorrelation function, $R(l)$, of height profile function $z(x)$ is integration of the product of the latter by its shifted form over the entire domain of x. Parameter l is the distance of shift along the x direction.



Average Reynolds Equation, Fig. 2 Asperity footprints of rough surfaces and corresponding Peklenik Number, γ

The autocorrelation function, $R(l)$, and its nondimensional form, $r(l)$, are defined below:

$$R(l) = \frac{1}{N - N_l} \sum_{1}^{N-N_l} (z_i(x) z_i(x+l)) = \lim_{L\to\infty} \frac{1}{L} \int_0^L z(x) z(x+l)\, dx \tag{3}$$

where L is the sampling length, and N_l the number of data within l.

$$r(l) = R(l)/R_q{}^2 \tag{4}$$

It is easily seen that when l=0, the autocorrelation function, R(l), becomes the RMS roughness and the nondimensional autocorrelation function, r(l), becomes 1.

The correlation length, l^*, measures the decay of a random event. It is the value of l over which r(l) drops to A% of its value at the origin (which is one, or the square of the RMS roughness when dimensional). Here, A% may be 10%, 20%, 37% (or 36.8%, which is 1/e), or 50%, all commonly seen in research papers. The correlation length may be used to determine the orientation of surface roughness. A parameter, the Peklenik number (Peklenik 1967), γ, or the asperity aspect ratio, is defined as the ratio of the correlation lengths in two orthogonal directions:

$$\gamma = l_x^*/l_y^* \tag{5}$$

The Peklenik number indicates surface texture orientation. γ equal to one, or around one, represents an isotropic surface, while a transverse surface can be defined by γ<1. The term *transverse* implies that the orientation of asperity "ridges" characterized by the x-direction correlation length is perpendicular to the x-direction, or the direction of the relative motion of the surfaces. Therefore, three typical surface orientations, isotropic, longitudinal, and transverse, can be characterized with γ=1, γ>1, and γ<1. Figure 2, based on Patir and Cheng (1978), illustrates the asperity footprints of such surfaces. Note that the footprints of longitudinal and transverse surfaces are similar. Rotating the footprints of a longitudinal roughness for about 90° results in those of a transverse roughness. Patir and Cheng (1978) used γ=9 as the required large number to represent a longitudinal surface. Later, Lee and Ren (1996) found that γ=6 should be sufficient for a clearly longitudinal orientation.

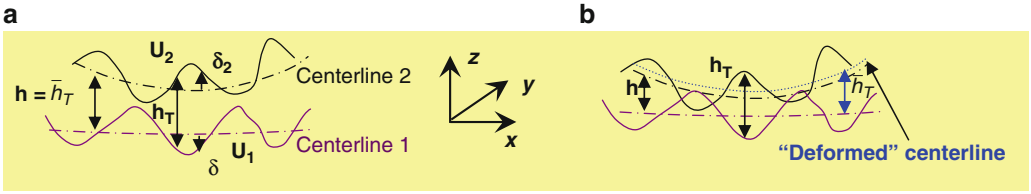
Average Flows

Figure 3 describes the lubrication of two rough surfaces. Note that asperities are exaggerated for clarity. The surface roughness here is a "Reynolds roughness," (i.e., the asperities have small slopes of a few degrees). Several terms are defined below:

U_1, U_2: Surface velocities
$\delta_1(x,y)$,$\delta_2(x,y)$: Surface deviation from the centerline
$h_T(x,y)$: Gap between two surfaces
$\bar{h}_T(x,y)$: Average gap
h(x,y): Compliance (film thickness, by Patir and Cheng), defined as the separation between two original centerlines.
$\bar{h}_T(x,y)$: Average gap, defined as the separation between two deformed centerlines
η: Viscosity of the lubricant

If there were no contact and deformation, the compliance and the average gap should remain the same. The compliance and the average gap are not the same once deformation is involved. The latter should be calculated with the deformed surface profile obtained from an elastic or elastoplastic analysis.

The concept of the average gap is defined on the separation of the centerlines of deformed surfaces. The Reynolds equation is valid for the deterministic description of the flow between two surfaces with



Average Reynolds Equation, Fig. 3 Lubrication film and interaction of two rough surfaces. (**a**) Original surface roughness before deformation (**b**) Surface 2 is deformed under loading

Reynolds roughness, and the corresponding film thickness is the gap at a position of interest.

$$\frac{\partial}{\partial x}\left(\frac{h_T^3}{\eta}\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{h_T^3}{\eta}\frac{\partial p}{\partial y}\right)=6(U_1+U_2)\frac{\partial h_T}{\partial x}+12\frac{\partial h_T}{\partial t} \tag{6}$$

The corresponding flow rates in the orthogonal x and y directions are

$$q_x=-\frac{h_T^3}{12\eta}\frac{\partial p}{\partial x}+\frac{U_1+U_2}{2}h_T \tag{7}$$

$$q_y=-\frac{h_T^3}{12\eta}\frac{\partial p}{\partial y} \tag{8}$$

Modifying these flows with flow factors φ leads to the terms of average flows:

$$\bar{q}_x=-\phi_x\frac{\bar{h}_T^3}{12\eta}\frac{\partial \bar{p}}{\partial x}+\frac{U_1+U_2}{2}\bar{h}_T+\frac{U_1-U_2}{2}R_q\phi_s \tag{9}$$

$$\bar{q}_y=-\phi_y\frac{\bar{h}_T^3}{12\eta}\frac{\partial \bar{p}}{\partial y} \tag{10}$$

Flow factors are introduced for flow modification, and an additional shear flow term is added to the flow in the entrainment direction. Note that the Patir and Cheng model defines the film thickness with respect to compliance, and the average flows are expressed as follows:

$$\bar{q}_x=-\phi_x\frac{h^3}{12\eta}\frac{\partial \bar{p}}{\partial x}+\frac{U_1+U_2}{2}\bar{h}_T+\frac{U_1-U_2}{2}R_q\phi_s \tag{11}$$

$$\bar{q}_y=-\phi_y\frac{h^3}{12\eta}\frac{\partial \bar{p}}{\partial y} \tag{12}$$

In Equations (9) through (12), φ_x and φ_y are pressure flow factors, and φ_s is the shear-flow factor, which are defined in the entry "► Flow Factors for Average Reynolds Equation." The additional term, $(U_1-U_2)R_q\varphi_s/2$, shows the flow due to surface roughness. For smooth surfaces, R_q is 0 and so is the additional flow. U_1 and U_2 have different influences on this flow. For $U_1=U_2$, no additional flow enters the region of interest.

Average Reynolds Equation

A differential control volume, $\Delta x\Delta y\bar{h}_T$, can be defined with the rough surfaces shown in Fig. 3 with $\bar{h}_T$ as the height of this volume. The flow balance results in the following:

$$\bar{h}_T\Delta y(\bar{q}_x+\frac{\partial \bar{q}_x}{\partial x}\Delta x)-\bar{h}_T\Delta y\bar{q}_x+\bar{h}_T\Delta x(\bar{q}_y+\frac{\partial \bar{q}_y}{\partial y}\Delta y)$$
$$-\bar{h}_T\Delta x\bar{q}_y+\frac{\partial \bar{h}_T}{\partial t}\Delta x\Delta y\bar{h}_T=0$$

which is

$$\frac{\partial \bar{q}_x}{\partial x}+\frac{\partial \bar{q}_y}{\partial y}=-\frac{\partial \bar{h}_T}{\partial t} \tag{13}$$

Substituting (11) and (12) to the equation above results in the average Reynolds equation:

$$\frac{\partial}{\partial x}\left(\phi_x\frac{h^3}{\eta}\frac{\partial \bar{p}}{\partial x}\right)+\frac{\partial}{\partial y}\left(\phi_y\frac{h^3}{\eta}\frac{\partial \bar{p}}{\partial y}\right)=6(U_1+U_2)\frac{\partial \bar{h}_T}{\partial x}$$
$$+6(U_1-U_2)R_q\frac{\partial \phi_s}{\partial x}+12\frac{\partial \bar{h}_T}{\partial t} \tag{14}$$

In the equation above, the flow factors are functions of film thickness and two statistical parameters, the RMS roughness and orientation parameter (the aspect ratio). Empirical formulas for flow factors for surfaces with a Gaussian asperity height distribution have been derived by Patir and Cheng (1978). However, one should realize that two statistical parameters might be insufficient to define an engineering surface. Also, because the flows are averaged across the film thickness, the average Reynolds equation does not allow cross-film property variations. Some strong assumptions (Newtonian fluid, no surface deformation, and no flow cavitation) were made in the derivation. Most importantly, mixed film terms, the average gap and compliance, are used in the equation

and, therefore, an additional relationship is needed to convert the average gap to the compliance, or vice versa.

Because compliance does not take into consideration surface deformation, a more reasonable solution to this problem may be to redefine the pressure flow factors in the equation shown above using the average gap given in (9) and (10). As a result, the pressure flow factor, φ_{Tx}, defined based on the average gap, becomes that shown in (15). Note that subscript T indicates average gap, the same as that in $\bar{h}_T$.

$$\phi_{gx} = \frac{\frac{1}{L_y}\int_0^{L_y} \left(\frac{h_T^3}{12\eta}\frac{\partial p}{\partial x}\right) dy}{\frac{\bar{h}_T{}^3}{12\eta}\frac{\partial \bar{p}}{\partial x}} \tag{15}$$

One can see the following relationship between φ_{Tx} defined based on average gap and the original Patir-Cheng flow factor, φ_x, defined on compliance (Shi and Wang 1998).

$$\phi_{Tx} = \frac{h^3 \phi_x}{\bar{h}_T{}^3}$$

The shear flow factor is not affected by the gap-compliance conversion. Note that viscosity may vary across the film due to the thermal effect or other non-Newtonian effects. The average Reynolds equation for the lubrication of rough surfaces with a non-Newtonian lubricant should be derived case-by-case based on the rheological model (Li et al. 1997).

Key Applications

Although only two statistical parameters are used and only orthogonal flow modifications are considered, the Patir-Cheng average Reynolds equation seems to be a reasonable simplification. Simplicity makes this equation convenient to use.

Interactive Stochastic-Deterministic Modeling of the Lubrication of Journal Bearing Conformal-Contact Problems

Conformal-contact lubrication problems involve the analyses for asperity interaction and structure deformation/deflection in a lubrication solution. Journal bearings are examples of such systems. Here, one has to deal with issues of at least two different scales: macro geometry and micro/nano scale asperities. An interactive deterministic-stochastic approach (or the macro-micro approach) (Shi and Wang 1998, Wang et al. 2004) may be a means to tackle such problems.

The film thickness given below incorporates two different scales. Here, surface elasticity may be divided into two parts: the deflection of the middle surface of asperities (surface formed by the middle points across roughness heights of a three-dimensional topography, similar to the centerline for a two-dimensional roughness), u_{zMS}, and the deformation of the asperities with respect to the middle surface, δ_1 and δ_2. The latter affects the micro flows and may be considered by an average flow (AF) model, such as the one by Patir and Cheng mentioned above, while the former is a part of the structural deformation and should be analyzed through a macro-scale elasticity computation (Wang et al. 2004).

$$h = h_0(t) + B_x x^2 + B_y y^2 + \underbrace{u_{zMS}(x,y,t)}_{\text{to be considered at the macro scale}} + \underbrace{\delta_1(x,y,t) + \delta_2(x,y,t)}_{\text{to be considered in the AF model}} \tag{16}$$

The macro-micro approach includes three components: (a) the macro middle-surface determination, u_{zMS}, (b) flow-factor and pressure calculation, and (c) micro contact analysis at each calculation location. An assumption has to be made in order to use the average-flow model; the distortion of asperities is negligibly small. The rough surfaces in interaction are superimposed to the film thickness determined by considering the global elasticity. Therefore, the average film thickness equation becomes

$$h = h_0(t) + B_x x^2 + B_y y^2 + u_{zCL}(x,y,t) \tag{17}$$

The average Reynolds equation is applied to each mesh grid where the film thickness can be treated as a constant. Flow factors are then determined corresponding to this film thickness. An off-line asperity contact model needs to accompany the analysis process if asperity contact is an issue. A pressure-gap function can be prepared off line to be retrieved by the lubrication procedure. A given film thickness (which is the gap in a contact simulation) corresponds to a level of asperity contact. Therefore, the contact pressure, $p_c(x,y)$, can be related to this off-line pressure-gap function as follows:

$$p_c(x,y) = f\left(\frac{h}{R_q}\right) \tag{18}$$

The geometry of the journal-bearing concave element system cannot be treated as an infinitely large body. Neither the Boussinesq nor the Flamant solution can be simply applied here. The finite element method can be utilized either for seeking the full deformation/deflection solution or for preparing influence functions to be used for surface deformation calculation.

Interactive Stochastic-Deterministic Modeling of the Lubrication of Counterformal Contact Problems

The macro-micro approach mentioned above has also been used to model the mixed elastohydrodynamic lubrication (EHL) problems in counterformal contacts**.** The work by Wang et al. (2004) is one of the examples of such efforts. With this approach, Patir and Cheng's average flow model can be employed to obtain the distribution of piecewise average pressure. The contact-embedment method, mentioned in the previous section, which incorporates details of asperity contact pressure into the overall pressure distribution, is utilized to reveal the severity of surface interaction. Wang et al. (2004) conducted numerical experiments, and the results were compared with those obtained by means of a full-scale numerical mixed-EHL model (Zhu and Hu 2001). The regime of the application of this macro-micro approach was also explored. The results of this work suggest that the macro-micro approach may yield reasonable film thickness and pressure in a local average sense for operating conditions corresponding to the ratio of the average film thickness to the RMS roughness, h_a/R_q, no less than 1. In this region, the film pressure may be approximated by the superposition of the hydrodynamic pressure and the asperity contact pressure. Most of the result disagreement appears in the regions where asperity interaction is predominately responsible for the load-supporting mechanism. As compared with the unified treatment of the hydrodynamic pressure and asperity contact in the full-scale mixed EHL model, superposition of the asperity contact pressure obtained from an off-line contact simulation to the hydrodynamic pressure from the average flow analysis under estimate asperity deformation but over estimate the average asperity pressure. Moreover, the accuracy of the macro-micro approach depends on the accuracy of flow factors.

Cross-References

► Elastohydrodynamic Lubrication (EHL)
► Mixed EHL
► Reynolds Equation

References

H. Christensen, Stochastic models for hydrodynamic lubrication of rough surfaces. Proc. Inst. Mech. Eng. 1969–1970 **184** (Part 1, 55), 1013–1026 (1970)

S. Harp, R. Salant, An average flow model of rough surface lubrication with inter-asperity cavitation. J. Tribol. **123**, 134–143 (2001)

S.C. Lee, N. Ren, Behavior of elastic-plastic rough surface contacts as affected by the surface topography, load and materials. STLE Tribol. Trans. **39**, 67–74 (1996)

N. Letalleur, F. Plouraboue, M. Prat, Average flow model of rough surface lubrication: Flow factors for sinusoidal surfaces. ASME J. Tribol. **124**, 539–546 (2002)

W.-L. Li, C.-I. Weng, C.-C. Hwang, An average Reynolds equation for non-Newtonian fluid with application to the lubrication of the magnetic head-disk interface. Tribol. Trans. **40**, 111–119 (1997)

N. Patir, H.S. Cheng, An average flow model for determine effects of three dimensional roughness on partial hydrodynamic lubrication. ASME J. Lubr. Technol. **100**, 12–17 (1978)

J. Peklenik, New development in surface characterization and measurement by means of radon process analysis. Proc. Inst. Mech. Eng. **182** (3K), 108 (1967)

F. Sahlin, *Lubrication, Contact Mechanics and Leakage between Rough Surfaces*, Ph. D. Thesis, Luleå University of Technology, Sweden, 2008

F. Shi, Q. Wang, A mixed-TEHD model for journal bearing conformal contacts, Part I: Model formulation and approximation of heat transfer considering asperity contacts. ASME J. Tribol. **120**, 198–205 (1998)

Q. Wang, D. Zhu, T. Yu, H.S. Cheng, J. Jiang, S. Liu, Mixed lubrication analyses by a micro-macro approach and a full-scale micro EHL model. ASME J. Tribol. **126**, 81–91 (2004)

D. Zhu, Y.Z. Hu, A computer program package for the prediction of EHL and mixed lubrication characteristics, friction, subsurface stresses and flash temperatures based on measured 3-D surface roughness. Tribol.Trans. **44**, 383–390 (2001)

Averaging of the Reynolds Equation

► Homogenization of the Reynolds Equation

Aviation Turbine Engine Oil Application

Susan C. Brown[1], Ronald E. Yungk[2]
[1]Consultant, retired Pratt & Whitney, Johnson City, NY, USA
[2]Air BP Lubricants, Naperville, IL, USA

Synonyms

Aircraft lubricants; Turbine engine lubrication; Turbine engine oils

Definition

Aviation lubricants encompass a wide class of lubricating oils used in aircraft propulsion systems, specifically to lubricate and cool the engine mechanical system. Primary mechanical system components that are typically lubricated with engine oil include engine mainshaft bearings, accessory drive gearbox components, and sump seals.

Introduction

The history of the gas turbine engine dates back to 1908, when French engineer Rene Lorin proposed the theory of gas turbine propulsion. Lorin proposed using a piston engine to compress air that would then be mixed with fuel and burned, which would produce pulses of hot gases that would then be expelled through a nozzle. Practical application of this theory was demonstrated in the late 1930s, with the inaugural flight on August 27, 1939, of the Heinkel He 178, powered by the HeS 3b gas turbine engine, developed in Germany by Dr. Hans von Ohain. Concurrently, in Britain, Sir Frank Whittle developed the Power Jet W.1 gas turbine engine. This engine first flew on the British Closter G.40 on May 15, 1941. By the end of World War II, both the von Ohain and Whittle engines were successfully powering fighter aircraft.

Although the axial-flow compressor design on the von Ohain engine differed from the centrifugal compressor design of the Whittle engine, the function of the mechanical system of both engines was essentially the same. And, although the performance requirements, i.e., speed, load-carrying capability, temperature capability, life, etc., of mechanical system components have changed over the years, the function of the mechanical system of modern gas turbine engines is essentially the same as it was for the original von Ohain and Whittle engines.

The gas turbine mechanical system is the engine rotor support system. Some of the primary components of the mechanical system may include bearing systems (rings, cages, rollers/balls, bearing supports, etc.), gears, seals, o-rings, gearbox, and the lubrication system (lubricant, oil supply and scavenge pumps, oil tank, filter, fuel/oil cooler, deoiler, deaerator, oil slingers, etc.). All of these mechanical system components must function properly for the mechanical system, and subsequently the engine, to function properly.

The gas turbine engine lubricant is an important component of the mechanical system. The primary functions of the lubricant are to lubricate (i.e., reduce friction and wear of bearings, gears, and other rotating components), cool the lubrication system components, and transport debris away from the lubrication system components and into the main oil filter. The lubricant chemistry must be compatible with all metallic and non-metallic lubrication system components. The evolution of aircraft turbine engine oils and critical properties and the impact of those properties on jet engine design are discussed below (Brown et al. 2001).

Turbine Engine Lubrication

Aircraft gas turbines developed prior to 1948 were lubricated satisfactorily with light mineral oils. As power requirements increased and engine cycle temperatures rose, mineral oils lacked the necessary high-temperature stability. Increasingly aggressive aviation turbine engine designs and the resulting oxidation and thermal degradation of the early lubricants led to a search for a new class of oils for aviation.

Synthetic ester base fluids were developed for aircraft turbine engine in the late 1940s and are still used today because of their inherently desirable characteristics. The most significant attribute of fully formulated synthetic esters is their good thermo-oxidative stability, as evidenced by high bulk fluid temperatures up to 200°C and thermal decomposition temperatures in transient hot spots up to 370°C. Synthetic esters also have very good temperature-viscosity response (viscosity index), which is necessary for the range of ambient to operating temperatures encountered in aviation. Viscosity index improvers cannot be used because these additives do not have the necessary thermo-oxidative stability to participate in this challenging formulation. The properties of aviation turbine oils are further enhanced by the addition of antioxidants, metal deactivators, load carrying agents, etc., which are miscible and stable in the relatively polar esters.

Ester molecules used as lubricating base oils are typically formed from alcohols with two to six hydroxyl functionalities. Organic acids of varying carbon chain length and structure are used to form ester base stock molecules with the desired viscosity.

With the selection of the proper base oil and the incorporation of the necessary additive systems, fully formulated ester oils have a broad spectrum of good properties including: (1) excellent deposit control, (2) good low-temperature fluidity, (3) good oxidative and thermal stability, (4) good bearing and gear fatigue resistance, (5) good gear load-carrying ability, (6) good corrosion protection, and (7) good compatibility with all engine materials of construction, especially non-metallic seals and soft metals. Among these, the major feature most important to modern turbine engine operation is good deposit control.

Many engines used in aircraft do not receive full overhauls until 20,000–40,000 h have elapsed, thus any deposits formed within the engine may cause lubrication system problems if not inspected and cleaned properly. Such problems are manifested by (1) plugging of oil jets, resulting in oil starvation to bearings and gears, (2) plugging of the breather system, resulting in excessive oil system breather compartment pressures leading to higher than normal compartment temperature and severe oil leakage, (3) plugging of scavenge systems, resulting in severe oil loss, and (4) fouling of carbon face shaft seals

rendering them inoperative allowing excessive air leakage into the oil system, resulting in excessively rapid oil degradation and deposit formation.

Classification of Turbine Engine Oils

Most of the turbine engine oils used today are described in military specifications MIL-PRF-23699 (MIL-PRF-23699 Performance Specification, "*Lubricating Oil, Aircraft Turbine Engine, Synthetic Base, NATO Code Number O-156*") and MIL-PRF-7808 (MIL-PRF-7808 Performance Specification, "*Lubricating Oil, Aircraft Turbine Engine, Synthetic Base*") (Table 1). The oil described by MIL-PRF-23699 is a 5 cSt viscosity grade (at 100°C) and serves not only U.S. Army and Navy turbine applications, but most commercial and general aviation turbine applications worldwide. There are now three classes of oil described in MIL-PRF-23699: STD or Standard class, HTS or High Thermal Stability class, and C/I or Corrosion Inhibiting class. The HTS class of oil uses very stable ester base-stocks and advanced antioxidant packages to achieve optimized thermo-oxidative stability for the purpose of minimizing deposits that limit time between overhauls (TBO). The C/I class of oil uses a corrosion inhibition package to minimize corrosion in military applications where operating environments are particularly corrosive.

MIL-PRF-7808 consists of two viscosity grades, 3 and 4 cSt (at 100°C) and is utilized in applications requiring low temperature start capability as low as −54°C. These applications include U.S. Air Force fighter aircraft in cold climates as well as commercial aircraft auxiliary power units (APUs) that endure cold temperatures at altitude. The more viscous MIL-PRF-7808, Grade 4 also has higher thermal stability requirements recognizing the need for more stable oils to for advanced military aircraft that operate at increasingly hotter temperatures. Development efforts between the U.S. Air Force, U.S. Navy, engine OEMs and oil manufacturers are expected to lead to an advanced 5 cSt oil with HTS class properties with enhanced load carrying capability in 2011.

So-called high load carrying oils are typically used for helicopter transmission gearboxes and some gas turbine systems with highly loaded gear sets. These 5 cSt grade oils are described in DOD-PRF-85734 (DOD-PRF-85734 Performance Specification, "*Lubricating Oil, Helicopter Transmission System, Synthetic Base*") are similar to the MIL-PRF-23699 oils with an additional Extreme Pressure (EP) additive system that promotes better boundary film lubrication in high load situations. Development efforts between the U.S. Navy, helicopter OEMs, and oil manufacturers are expected to lead to an 9 cSt Advanced Helicopter Transmission Lubricant in 2010.

Aviation Turbine Engine Oil Application, Table 1 Grades and classes of turbine engine oil

Specification	Viscosity cSt at 100°C	Attributes
MIL-PRF-7808, grade 3[b]	3	Lowest temperature start capability
MIL-PRF-7808, grade 4[b]	4	Improved cleanliness/ stability with good low temperature start capability
MIL-PRF-23699, class STD[a]	5	Higher viscosity for improved system durability
MIL-PRF-23699, class HTS[a]	5	Best cleanliness/stability
MIL-PRF-23699, class C/I[a]	5	Corrosion inhibition for salt environments
DOD-PRF-85734[c]	5	High load carrying for helicopter gearboxes

[a]MIL-PRF-23699 Performance Specification, *"Lubricating Oil, Aircraft Turbine Engine, Synthetic Base, NATO Code Number O-156."*
[b]MIL-PRF-7808 Performance Specification, *"Lubricating Oil, Aircraft Turbine Engine, Synthetic Base."*
[c]DOD-PRF-85734 Performance Specification, *"Lubricating Oil, Helicopter Transmission System, Synthetic Base."*

The forum for many turbine oil issues, performance evaluation methods and specification development is an SAE Standards Development committee, E-34 (Propulsion Lubrication). E-34 published specification AS5780A (SAE Aerospace Standard AS5780A (or later revision), "Specification for Aero and Aero-Derived Gas Turbine Engine Lubricants") for 5 cSt turbine oils used in commercial aviation and aero-derived industrial and marine applications. OEM representatives on this committee form a Qualified Product Group that qualifies oils to AS5780A and manages change associated with those approved products. These OEMs use this specification as part of the aviation authority regulated commercial engine oil approval activities. References to AS5780A are increasingly replacing or augmenting the references to MIL-PRF-23699 in aircraft, engine, and component maintenance manuals.

Important Oil Properties and Test Methods

Aviation turbine engine oils are evaluated for lubricant attributes appropriate to their base chemistry, engine materials of construction, flight envelope parameters, and hardware time between overhauls.

Viscosity (ASTM D 445) – Controls lubrication film thickness and helps define the operating temperature range of the turbine lubrication system. Pressure – viscosity coefficient is also important to designers to accurately calculate lubrication film thickness.

Pour point (ASTM D 97) – Relates to the ability to service and start the engine in extremely cold weather.

Flash point (ASTM D 92) – Used to control flammability and volatility of the oil. A full understanding of flammability requires insight into many system parameters (residence time, fuel/air ratio, auto-ignition temperature).

Total Acid Number (ASTM D 664) – Acidic material in an oil may corrode bearings and valve guides and must be controlled.

Lubricant compatibility (FED-STD-791, Method 3403) – Evaluates the hot aging miscibility of candidate oils with other established oils likely to be encountered in service. This is important given the possible additive package diversity as allowed by performance (not material) specifications.

Elastomer compatibility (e.g., Def Stan 05–50 (Part 61) Method 22) – Measures swell and deterioration of seal materials in contact with hot oil. Elastomers of interest include fluorocarbon, perfluorocarbon, fluorosilicone, silicone, and nitrile. This becomes important because some high thermal stability ester base stocks and anti-oxidant packages tend to be aggressive toward fluorocarbon, while some extreme pressure additives are aggressive toward silicone and fluorosilicone.

Oxidation and Corrosion Stability (FED-STD-791, Method 5308 mod 1 or ASTM D 4636) – Evaluates bulk oil stability at temperatures up to 218°C as well as any breakdown products' consequential corrosive impact on representative metallurgies.

Thermal Stability and Corrosivity (FED-STD-791, Method 3411) – Quality control method for detecting undesirable contaminations from other non-aviation ester based products.

Deposit control – Measured in a full-scale, heated bearing rig (FED-STD-791, Method 3410, severity rating of 1.5). This test is run for 100 h (HTS oils require 200 h) with evaluation of oil condition control and a cleanliness demerit rating. Emerging subscale tests are being developed and evaluated to determine control of liquid phase deposits (SAE ARP 5996 to simulate oil pressure pipes and jets), mixed phase deposits (GE Alcor High Temperature Deposition Test to simulate scavenge system pipes) and vapor phase deposits (U.S. Navy test to simulate breather system pipes). This activity is indicative of the importance of controlling deposits in engines with ever-increasing time between overhauls.

Load carrying capability or boundary film lubricating ability – Measured by the Ryder Gear test (FED-STD-791, Method 6508) and/or Wedeven Associates Load Carrying Capability test (SAE AIR 4978 Appendix E). This property is especially important for the DOD-PRF-85734 oils that are also used in highly loaded helicopter transmission gearboxes.

Foaming tendency – Measured by either static (e.g., ASTM D892) or dynamic means (e.g., FED-STD-791, Method 3214). Oil can foam when subjected to the shearing forces of bearings and gears, especially at altitude or when influenced by impurities such as rogue silicone compounds.

Sediment or particulate contamination (FED-STD-791, Method 3010 or 3013). – The low maximum ash content allowed in MIL-PRF-23699 or AS5780A oils effectively precludes the use of organo-metallic additives.

Hydrolytic stability (Def Stan 05–50 (part 61) Method 6) – Important for storage stability and the stability of oils in capped (not breathed) lubrication systems typical of the designs used in engine accessories such as generators and air starters. Hydrolysis is a degradation chemical reaction between an ester and water at elevated temperatures that reverts the ester to its original acid and alcohol destroying the turbine oil's properties.

Evaporation loss (ASTM D972) – Helps ensure lubrication systems do not consume too much oil during operation.

Shear stability – Controlled by ASTM D2603 ensuring oil products do not lose viscosity due to the mechanical forces encountered in aviation turbine engines.

Corrosion Inhibition – In MIL-PRF-23699, Class C/I oils is controlled by SAE ARP 4249 ball corrosion testing. This attribute of the C/I class oils is important for marine applications of aviation turbine oils.

Storage stability tests – Ensures all oil components remain miscible at low temperature and during extended durations up to 3 years as described in the military oil specifications.

Acid assay (FED-STD-791, Method 3500 (1)) – Controls the acid distribution used in an oil formulation's ester base-stock. This ensures the oil formulation is controlled on a batch-to-batch basis similar to the originally qualified formulation.

Trace metals – Controlled in new oils in order to provide a baseline for spectrographic oil analysis program diagnostic testing discussed below.

Spectrographic Oil Analysis Program (SOAP)

The parts per million (ppm) of several metals (e.g., Fe, Ag, Cr, Al, Mg, Ti, Mo, V) in a sample of used oil is usually determined by exposure of a sample to an emission

spectrometer. Spectrographic oil analysis was first used in the 1940s by the railroads to determine bearing wear in diesel engines, and was very successful. The military studied oil analysis for predicting power plant failure in aircraft engines and the practice is now in widespread use. Many analytical laboratories offer such service to airlines and operators of general aviation.

Provided the proper techniques are utilized in obtaining oil samples and sufficient background information is provided to the analytical laboratory on the type of engine and oil history, spectrometric analysis is a very useful tool in determining the condition of the oil. Caution should be observed in the use of one random sample of oil to assess any engine condition. Each engine has its own normal wear pattern; therefore trend monitoring with an understanding of a particular engine's wear pattern over time is the most meaningful approach to oil analysis. If an alarm is raised because of excessive wear metals in the spectrometric analysis of the oil, the typical practice is to immediately obtain a second confirmation sample of oil for a recheck before maintenance action is specified.

Many operators use a combination of engine parameter trending (oil consumption, temperature, pressure, etc.), periodic examination of magnetic chip collector, examination of oil filter/screen for wear debris, and/or spectrometric oil analysis to judge the condition of their engines. Magnetic chip collectors incorporated into the lubrication scavenge system also serve as a valuable diagnostic compliment for engine condition monitoring. These chip collectors are inspected by mechanics at regular ("letter" check) intervals. The debris collected can be evaluated with scanning electron microscopes equipped with energy dispersive spectroscopy (SEM/EDS) to determine the material type and morphology of the debris. This resulting information can often be interpreted to guide the maintenance actions.

Advanced lubrication system designs are beginning to incorporate oil and debris monitors that are integrated into the engine's control system providing a more comprehensive Prognostic Health Monitoring (PHM) system.

Cross-References

► O-rings
► Pour Point
► Scanning Electron Microscopy (SEM)

References

S.C. Brown, H.A. Chin, D.A Haluck, L.D. Wedeven, Linking Lubricants, Materials, Design & Tribology – Changes for Development of Advanced Mechanical Systems. Lubr. Fluid Power **2**(4), 7–21 (2001)

AW – Glass-Ceramic Apatite–Wollastonite

► Modified UHMWPE for the Hip Joint (Particle Filled and Reinforced)

Axial Bearings

► Thrust Bearings

Axial Bearings for Hydrogenerators

► Large Hydrodynamic Thrust Bearings and Their Application in Hydrogenerators

Axial Instability

► Self-Excited Gas Bearing Instabilities

Axial Thrust Bearings

► Thrust Bearings